

# 教育人工智能中的算法偏见： 生成机制与治理框架

丁勇  
潍坊科技学院

**摘要** 随着人工智能在教育领域的广泛应用，算法偏见逐渐成为影响教育公平的重要问题。本文以教育人工智能（AI in Education, AIEd）为研究对象，从社会—技术系统视角系统分析算法偏见的生成机制与治理路径。研究指出，教育 AI 中的算法偏见并非单纯的技术缺陷，而是结构性不平等在数据采集、模型训练、系统部署与人机交互等环节中的累积性再生产过程。文章从数据偏见、模型偏见、部署偏见与交互偏见四个维度构建算法偏见的全链路分析框架，并结合算法公平性理论，讨论群体公平、个体公平与因果公平之间的内在张力。研究进一步提出以“发展性公平”为核心理念，将公平理解为面向学习者长期发展的动态过程，而非单一时点的统计均衡。在此基础上，本文构建由数据审计、模型审计、影响评估与制度规范构成的多维治理框架，以期教育人工智能的规范发展与教育公平治理提供理论参考。

**关键词** 教育人工智能；算法偏见；发展性公平；算法公平；教育 AI 治理

DOI <https://doi.org/10.6914/aiesc.020103> 文章编号 3079-7918.2026.0201.84-108

## 引言

人工智能（AI）正以前所未有的速度和广度渗透进全球教育系统的每一个毛孔。从个性化学习路径推荐引擎到智能辅导系统（Intelligent Tutoring Systems, ITS），从自动化论文评分器到教师绩效增值模型（Value-Added Model, VAM），从入学录取预测算法到学生心理健康早期预警系统——人工智能教育（AI in Education, AIEd）应用正在以“提高效率”“精准施教”“数据驱动”的强势话语重塑着教育实践的基本形态<sup>①</sup>。据联合国教科文组织 2024 年发布的全球教育人工智能追踪报告，全球已有超过 170 个国家和地区在不同层级的教育系统中部署了某种形式的 AI 应用，其中学习分析（Learning Analytics）平台的年增长率高达 38%，生成式 AI 辅助教学工具在 2023—2024 学年间的渗透率从 4% 飙升至 27%<sup>②</sup>。这场技术革

**作者简介** 丁勇，潍坊科技学院教授，Email: ddy2004@126.com, <https://orcid.org/0000-0001-8225-1839>。

© The Author(s) 2026. CC BY-NC-ND 4.0. **Article History:** Received: 15 January 2026; Accepted: 20 February 2026; Published: 15 March 2026 (online), 15 March 2026 (print).

**人工智能教育研究** (Réngōng zhīnéng jiàoyù yánjiū) *Artificial Intelligence Education Studies*, ISSN 3079-7918 (Print), ISSN 3079-7926 (Online). Published by Creative Publishing Co., Limited 开元出版有限公司 (CPCL®). Email: cpcl@cpcl.hk, <https://aies.press>, <https://cpcl.hk>.

<sup>①</sup>Holmes W, Bialik M, Fadel C. *Artificial Intelligence in Education: Promises and Implications for Teaching and Learning* [M]. Boston: Center for Curriculum Redesign, 2019.

<sup>②</sup>UNESCO. *Global Education Monitoring Report 2023: Technology in Education* [R]. Paris: UNESCO Publishing, 2023.

命的规模之大、速度之快、渗透之深，在人类教育史上绝无先例。

然而，在这片看似光芒万丈的技术乐土之下，一道深不见底的裂隙正在悄然扩张。2020年，英国 A-level 考试因新冠疫情取消统考后，英格兰考试监管机构 Ofqual 启用算法模型为学生派定成绩。该算法以学校历史表现数据为核心权重指标，本意是确保分数分布的“统计稳定性”。然而，其运行结果令全社会震惊：算法系统性地将公立学校和弱势社区学生的分数大幅压低——在某些案例中降幅超过两个等级——而私立学校学生则普遍获得了与教师预测一致甚至更高的分数<sup>①</sup>。究其原因，私立学校因班级规模小、历史成绩波动大而被算法赋予了更高的“不确定性容忍度”，直接采用了教师预测分数；而公立学校因学生基数庞大、历史数据“稳定”而被算法牢牢锁定在其既往的统计轨道上——换言之，那些历史上就因资源匮乏而成绩偏低的学校，被算法判定为“本来就应该这么低”。这一“算法丑闻”迫使英国政府在舆论的燎原之火中紧急撤回了全部算法成绩，影响近 50 万名学生的人生轨迹<sup>②</sup>，成为全球教育 AI 治理史上最具标志性的警示事件。它以血淋淋的方式揭示了一个核心命题：算法偏见（Algorithmic Bias）不是一个边缘的技术瑕疵，而是结构性不平等（Structural Inequality）在数据—模型—部署全链路中的系统性映射与放大<sup>③</sup>。

英国 A-level 事件绝非孤例。在大西洋彼岸，美国多个州的公立学区长期使用教师增值模型（VAM）来评估教师绩效——该模型通过比较学生在某位教师指导下的实际成绩与系统预测的“期望成绩”之间的差异来衡量教师的“附加价值”。然而，多项独立研究发现，VAM 对教师的评级高度不稳定（同一教师在不同年份的评级波动可达四个等级），且系统性地惩罚那些执教于高贫困、高流动性学区的教师——不是因为这些教师教得差，而是因为他们面对的学生群体的基线本身就更具挑战性<sup>④</sup>。一些杰出教师因为 VAM 的低分评级而被解雇或降薪，暴露了算法在面对复杂教育情境时将“结构性劣势”错误归因为“个人能力缺陷”的根本性谬误。

在中国，算法偏见同样在教育实践中投下了浓重的阴影。随着“智慧教育”战略的大力推进，各类 AI 教育产品如雨后春笋般涌现——智能组卷系统、自适应学习平台以及 AI 驱动的考生分析与学情诊断工具已覆盖了数以千万计的中小学生和大学生<sup>⑤</sup>。然而，这些系统的训练数据高度集中于东部发达地区的优质学校，导致其在面对西部农村学校、少数民族地区学校以及特殊教育群体时表现出显著的预测偏差<sup>⑥</sup>。一个在北京市海淀区重点中学数据上训练出来的学情诊断系统，当被移植到云南省怒江州的乡村小学时，其诊断准确率可能从 92% 骤降至不足 60%——不是因为算法本身有“bug”，而是因为算法从未“见过”这类学生的数据。这种“数据盲区”造成的偏见，在技术开发者的认知地图中往往是完全隐形的。

更值得警醒的是，随着 2023 年以来 ChatGPT、Claude、Gemini 等生成式大语言模型（LLM）在教育领域的爆发式渗透，算法偏见的形态正在经历一场深刻的“范式跃迁”。传统的 AIED 算法偏见主要是分类预测型的——即模型对不同群体给出差异化的预测标签（如“高风险/低风险”“通过/不通过”），其偏见相对容易通过群体公平性指标来检测。然而，生成式 AI 的偏见嵌入方式发生了根本性变化：它不再是给出一个离散的标签，而是生成整段连续的自然语言文本——课程大纲、教学建议、学生评语、论文反馈、考试题目——而这些文本中潜藏的偏见远比一个预测标签更加隐蔽、更加微妙、也更加难以测量<sup>⑦</sup>。当一个 LLM 为一名女性学数学的学生生成的鼓励性反馈在语气、深度和期望水平上系统性地低于为男性同伴生

①Meadows M, Black B, Sherlin G. Ofqual's Approach to Grading in Summer 2020 [R]. Coventry: Ofqual, 2020.

②Kolkman D. The (in)Credibility of Algorithmic Models to Non-Experts [J]. *Big Data & Society*, 2021, 8(1): 1-14.

③Baker R S, Hawn A. Algorithmic Bias in Education [J]. *International Journal of Artificial Intelligence in Education*, 2022, 32(4): 1052-1092.

④Amrein-Beardsley A. Rethinking Value-Added Models in Education: Critical Perspectives on Tests and Assessment-Based Accountability [M]. New York: Routledge, 2014.

⑤祝智庭, 韩中美, 黄昌勤. 教育人工智能 (eAI): 人本人工智能的新范式 [J]. 电化教育研究, 2021(1): 5-15.

⑥中国信息通信研究院. 人工智能发展报告 (2024 年)[R]. 北京: 中国信息通信研究院, 2024.

⑦Kasnezi E, Seßler K, Küchemann S, et al. ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education [J]. *Learning and Individual Differences*, 2023, 103: 102274.

成的反馈时,这种差异几乎不会被任何现有的公平性度量指标所捕获——因为两段文字表面上都看似“正面”和“鼓励”。这种隐性偏见(Latent Bias)构成了AIED算法治理的全新前线。

教育领域的算法偏见之所以格外危险,还有一个更为根本的原因:教育本身承载着社会流动、阶层跃迁与代际公平的双重使命<sup>①</sup>。在人类社会的制度体系中,教育是为数不多的能够打破阶层固化、实现向上流动的“梯子”。而当一个算法系统“学习”了过去数十年间嵌入社会结构的种族、性别、阶层偏见后,它不是中性地“预测”未来——它是在以数学的精准度“复制”过去的的天不公,并赋予这种不公以“客观”“科学”的外衣<sup>②</sup>。更为险恶的是,算法偏见在教育中具有独特的“反馈循环效应”(Feedback Loop):当系统预测某位学生“不太可能成功”并据此减少对该生的资源投入时,该生的表现果然下降,而这一下降又反过来“验证”了算法最初的偏见性预测<sup>③</sup>。这种自我实现的预言(Self-Fulfilling Prophecy)在教育领域的杀伤力尤为巨大——它不仅扭曲了个体学生的发展轨迹,更在宏观层面将偏见从可修正的错误转变为不可逆的系统性伤害,使得“梯子”变成了“天花板”,将本应承担社会公平器功能的教育系统异化为不平等的固化器。

面对这一严峻挑战,当前学界的主流应对策略仍然被困在“技术解决技术问题”的封闭回路中。大量研究聚焦于算法层面的“去偏见”(Debiasing)技术——数据重采样(Resampling)、模型约束优化(Constrained Optimization)、后处理校准(Post-hoc Calibration)等<sup>④</sup>。这些技术手段当然有其不可否认的工程价值:它们能够在特定的、定义清晰的公平性目标下显著降低模型输出的群体间差异。然而,本文将在后续章节中系统论证,仅靠“修复模型”来治理教育算法偏见,就像仅靠给一栋地基倾斜的建筑粉刷外墙来“解决”其倒塌风险一样——它遮蔽了问题的结构性根源,并以虚假的“技术修复叙事”延误了真正的制度性变革。更为深刻的是,去偏见技术本身面临一个数学上已被证明的根本困境:多种公平性度量指标在一般条件下不可能被同时满足<sup>⑤</sup>。这意味着“什么是公平”在数学上没有唯一解——它本质上是一个需要在特定社会-文化-政治情境中通过多方协商来界定的价值判断问题,而非一个可以由工程师在代码中单方面给出答案的技术参数优化问题。

这一核心洞察构成了本文批判性立场的知识论基础。在既有研究版图中,西方学界在算法公平性(Algorithmic Fairness)、可问责性(Accountability)与透明度(Transparency)——即所谓的“FAT原则”——方面已积累了丰富的理论成果<sup>⑥</sup>。然而,这些成果存在两个显著的盲区。第一,绝大多数研究以刑事司法(如预测性警务、累犯风险评估)和金融信贷(如贷款审批)为应用场景,直接面向教育领域的系统性研究仍然匮乏——而教育场景因其对发展性(学生处于成长之中)、可塑性(行为可以通过干预改变)和价值多元性(教育目标远不止于“效率”)的独特要求,与司法和金融领域存在本质差异,不可简单移植。第二,现有研究几乎全部在“北半球”(Global North)的语境中展开,对中国、印度等发展中大国教育系统中AI偏见的独特生态——如城乡二元数据鸿沟、多民族多语言数据生态、应试文化与素质教育的张力等——缺乏关照。

在中文学术界,关于AI教育偏见的研究正在起步但尚未形成系统性的理论体系。刘晋等学者通过对

①Williamson B. Big Data in Education: The Digital Future of Learning, Policy and Practice [M]. London: SAGE Publications, 2017.

②O'Neil C. Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy [M]. New York: Crown Publishers, 2016.

③Ensign D, Friedler S A, Neville S, et al. Runaway Feedback Loops in Predictive Policing [C]. Proceedings of the Conference on Fairness, Accountability, and Transparency. New York: ACM, 2018: 160–171.

④Mehrabi N, Morstatter F, Saxena N, et al. A Survey on Bias and Fairness in Machine Learning [J]. ACM Computing Surveys, 2021, 54(6): 1–35.

⑤Chouldechova A. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments [J]. Big Data, 2017, 5(2): 153–163.

⑥Selbst A D, Boyd D, Friedler S A, et al. Fairness and Abstraction in Sociotechnical Systems [C]. Proceedings of the Conference on Fairness, Accountability, and Transparency. New York: ACM, 2019: 59–68.

十项元分析的伞型评价揭示了数据驱动教学决策的效果异质性和视野局限<sup>①</sup>；中国科学院课题组在 2024 年 AI 发展报告中首次专辟章节讨论了教育 AI 的伦理风险；教育部在《教育信息化 2.0 行动计划》中也提出了“安全可控、伦理合规”的原则。然而，从“原则宣示”到“操作框架”之间，仍然存在一段巨大的鸿沟——需要的不是更多的口号，而是一套有理论深度、有技术精度、有制度支撑的系统性治理方案。

正是基于这一判断，本文提出“教育发展性公平”（Developmental Equity in Education）这一核心概念。本文的核心论点是：教育 AI 中的算法偏见不是一般意义上的技术偏差或模型缺陷，而是教育不平等的自动化再生产机制。与刑事司法或金融信贷领域不同，教育场景有其不可化约的特殊性：学生不是被“判断”的静态客体，而是正在成长中的发展性主体；教育决策不是一次性的终局裁决，而是沿着教育管道逐层累积的过程性选择。因此，教育 AI 的公平性不能照搬“静态公平”框架——它需要一种关注学习者发展权、承认教育过程可塑性、回应偏见累积效应的“发展性公平”范式。

基于这一核心命题，本文循“认识—治理—反思”三重递进逻辑展开。首先，从生成机制的考古学角度完整解剖算法偏见在教育 AI 系统中从数据到模型到部署的全生命周期——不仅追问“偏见长什么样”，更追问“偏见从哪里来”以及“偏见为什么在教育中特别致命”。其次，超越单维度的“去偏见”修补，在公平性度量指标的内在矛盾和“不可能定理”的理论背景下，构建一套融合技术审计、制度规范与利益相关者协商的多维审计框架。再次，系统梳理当前主流的去偏见技术路线（预处理—训练中—后处理），在承认其工程价值的同时严格界定其能力边界。最后，从技术哲学的高度批判“技术至上主义”（Techno-Solutionism）的意识形态陷阱，倡导从“修复模型”向“人机协同治理”的范式转型，并结合欧盟 AI 法案、英国 ATRS 标准与中国《生成式人工智能服务管理暂行办法》的全球比较视野，提出面向中国教育场景的本土化治理方案——一个使偏见能够被持续发现、被公开审议、被民主问责、被迭代修正的制度生态。

## 一、算法偏见的生成机制：全链路的系统性考古

算法偏见并非凭空出现在 AIEd 系统输出端的“飞来之祸”，而是在数据采集、模型训练、系统部署与用户交互的每一个环节中逐步累积、叠加和放大的系统性产物。要真正理解教育 AI 中偏见的本质，必须采取一种类似于考古学的“地层解剖”方法——不是仅仅观察地表的裂缝（即模型输出的群体间差异），而是一层一层地向下挖掘，追溯偏见在数据层、模型层、部署层和交互层中是如何被生成、被编码、被放大并最终被“自然化”的。本节从四个维度对这一全链路过程进行系统性还原。

### （一）数据偏见：历史不公的数字化冻结

数据偏见是算法偏见最为直接和普遍的源头，也是整个偏见生命周期中最容易被直觉理解、却最难以彻底根除的一环。其核心机制可以用一句话概括：AI 模型的训练数据本身就是特定历史-社会-文化条件下的产物，当这些数据被不加批判地输入模型时，嵌入其中的结构性不平等便以数学参数的形式被“冻结”在算法之中<sup>②</sup>。这种“冻结”的隐蔽性在于：数据在表面上看来只是一系列“客观”的数字记录——考试分数、出勤率、作业完成时间、课堂参与频次——但每一个数字的背后都潜伏着它生成时的社会条件。

数据偏见主要表现为三种相互交织的亚型。第一种是代表性偏见（Representation Bias）：训练数据集未能均匀覆盖所有目标群体，导致模型对“数据丰富”群体的预测精度远高于对“数据稀缺”群体的预测精度。这种偏见在教育领域的表现尤为突出。例如，语音识别系统的训练数据如果严重缺乏带有地方口音（如中国西南地区少数民族口音，或美国南方非裔方言 AAVE）的语料，该系统就无法准确转录这些学生的

<sup>①</sup>刘晋，张学波，林书兵. 数据驱动的教学决策效果总览与视野突破——基于十项元分析的伞型评价 [J]. 电化教育研究, 2022(9):53-62.

<sup>②</sup>Suresh H, Guttat J. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle [C]. Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization. New York: ACM, 2021: 1-9.

课堂发言<sup>①</sup>。Koencke 等人 2020 年在《美国国家科学院院刊》上发表的研究发现, 主流商用语音识别系统对非裔美国人英语的错误率是白人英语的近两倍。在教育场景中, 这意味着使用 AI 进行课堂讨论自动转录和语义分析时, 来自特定语言背景的学生的发言将被系统性地“听不懂”或“听错”, 进而导致其课堂参与度被低估。类似地, 面部识别系统若主要使用白人面孔数据集训练, 则对深色皮肤群体的识别准确率将急剧下降<sup>②</sup>。Buolamwini 和 Gebru 在其开创性的“Gender Shades”研究中揭示, IBM 和微软的面部识别系统对深色皮肤女性的失误率高达 34.7%, 而对浅色皮肤男性的失误率仅为 0.8%——这不是一个微小的性能差距, 而是一个数量级的鸿沟。当这些系统被部署到校园安全监控或考试身份验证等教育场景时, 特定群体的学生将面临系统性的识别失败和由此引发的差别化对待。

第二种是历史偏见 (Historical Bias): 即便训练数据在统计上“完美代表”了当前人口的构成比例, 但如果当前社会本身就存在深刻的结构性不平等, 那么这些数据忠实地“编码”的恰恰就是这种不平等。历史偏见是数据偏见中最为根深蒂固的形态, 因为它不是一个可以通过“多收集一些数据”来解决的采样问题, 而是数据“如实反映了一个不公平的世界”所天然携带的先天缺陷。在教育领域, 历史偏见的典型案例触目皆是: 使用过去二十年中国高考录取数据去训练一个“专业适配度预测模型”, 由于历史上工科专业女性录取率远低于男性 (这是性别社会化和结构性歧视的产物, 而非女性理工能力不足的证据), 模型将不可避免地“学会”对女性申请者给出更低的适配度评分。类似地, 使用过去十年性别失衡的雇用数据去训练招聘筛选模型, 模型自然也会“学会”歧视女性候选人——正如亚马逊公司著名的 AI 招聘工具所雄辩地揭示的<sup>③</sup>。在这些案例中, 算法没有“犯错”——它无比精确地完成了它被设计去完成的任务: 从数据中发现模式并据此预测。问题在于数据中的“模式”本身就是不公平的社会结构的高保真映射。

第三种是测量偏见 (Measurement Bias): 这是教育领域最常被忽视却极为关键的偏见类型。测量偏见指的是用来“捕获”某一概念 (如“学业能力”“学习投入”“教师效能”) 的具体指标本身就存在对某些群体的系统性不利。例如, 以“在线学习平台的点击次数”作为“学习投入度”的代理测量, 将系统性地低估那些因家庭经济条件限制而无法获得稳定网络接入的农村学生的投入程度<sup>④</sup>。同样, 以“标准化测试成绩”作为“学术潜力”的唯一量化维度, 将系统性地遮蔽那些在创造力、领导力、动手能力等非标准化维度上表现卓越的学生——而后者往往恰恰集中在社会经济地位较低的群体中。测量偏见的阴险之处在于, 它在数据被采集之前就已经决定了哪些信息被“看见”、哪些被“遗漏”——这是一种发生在数据起点处的视域剪裁。

## (二) 模型偏见: 算法构造中的价值嵌入

如果数据偏见是“原材料”层面的污染, 那么模型偏见则是“加工工序”层面的二次扭曲。偏见不仅源自数据, 也源自算法设计、模型架构选择、优化目标设定以及正则化策略等一系列看似“纯技术”的工程决策之中。这一层面的偏见更为隐蔽, 因为它往往深藏于数学公式和代码逻辑的深处, 被技术的“中立性”话语层层包裹。

代理偏见 (Proxy Bias) 是模型偏见最典型也最具欺骗性的表现形式。当法律或政策明确禁止在决策中使用种族、性别、宗教等敏感属性时, 算法工程师会将这些变量从特征集中移除, 以为如此便可一劳永逸地消除歧视。然而, 算法是精于“侦探”的: 它会自动发现那些与敏感属性高度相关的“代理变量” (Proxy

<sup>①</sup>Koencke A, Nam A, Lake E, et al. Racial Disparities in Automated Speech Recognition [J]. *Proceedings of the National Academy of Sciences*, 2020, 117(14): 7684-7689.

<sup>②</sup>Buolamwini J, Gebru T. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification [C]. *Proceedings of the Conference on Fairness, Accountability and Transparency*. 2018: 77-91.

<sup>③</sup>Dastin J. Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women [EB/OL]. (2018-10-10)[2026-04-16]. <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G>.

<sup>④</sup>Jacobs A Z, Wallach H. Measurement and Fairness [C]. *Proceedings of the Conference on Fairness, Accountability, and Transparency*. New York: ACM, 2021: 375-385.

Variables), 并通过这些间接通道将被明文禁止的歧视因子重新引入决策模型<sup>①</sup>。在教育场景中, 邮政编码可以作为种族和社会经济地位的代理, 姓名拼写方式可以作为民族和国籍的代理, 家庭中第一代大学生身份可以作为文化资本匮乏的代理, 课外活动参与类型(马术 vs. 篮球)可以作为阶层身份的代理。一个在表面上“完全不知道”学生种族信息的录取算法, 仅凭邮政编码和课外活动数据就能以极高的准确率“推断”出学生的族裔背景——然后基于这种推断做出差别化决策。这揭示了一个令人不安的事实: 在高维特征空间中, “移除敏感属性”只是一种自欺欺人的鸵鸟策略<sup>②</sup>。

模型的不透明性(Opacity)构成了模型偏见的第二重屏障。当代 AIED 系统越来越多地采用深度神经网络(DNN)和大语言模型(LLM)等高维、非线性、端到端的模型架构。这些模型以数百万乃至数十亿个参数的“黑箱”形式给出预测结果, 没有人——包括模型的开发者自己——能够完整地解释模型何以做出某一特定判断<sup>③</sup>。Cynthia Rudin 在《自然·机器智能》上发表的标志性论文中尖锐地指出: 在高风险决策领域(包括教育), 应当直接使用可解释模型而非事后对黑箱模型进行“解释性装修”。这种不可解释性使得偏见的检测与纠正变得极为困难: 如果我们都不知道模型“为什么”做出某个判断——将某生标记为“高辍学风险”或将某教师评定为“低效能”——我们又如何判定该判断是否公平? 不透明性不仅是一个技术局限, 更在权力维度上制造了深刻的不对称: 模型的开发者(通常是商业公司的技术团队)垄断了对系统逻辑的理解权, 而受模型决策影响最深的主体——学生、教师、家长——被彻底排除在质疑和审查的可能性之外。

优化目标偏见(Objective Function Bias)是模型偏见中最为隐匿的第三种形态。每一个机器学习模型在训练时都需要一个明确的优化目标(损失函数), 而这个目标的选择本身就蕴含着深刻的价值取向。当一个 AIED 系统将“预测学业成绩”作为唯一的优化目标时, 它就在无形中将教育的全部意义压缩为一个可以用数字衡量的绩效指标——考试分数。批判性思维的培育、创造力的激发、公民意识的觉醒、情感的健康发展——这些教育的核心使命因为“不可量化”而在模型的优化视野中彻底消失<sup>④</sup>。Gert Biesta 将这种现象称为“可测量之暴政”(Tyranny of the Measurable): 当我们只优化可以测量的东西时, 我们事实上是在让测量工具来定义什么是“好的教育”——这是一种深层的价值倒置。

### (三) 部署偏见: 从实验室到田野的“语境蒸发”

一个在实验室条件下通过了所有公平性检测的模型, 在被部署到真实教育场景时仍然可能产生严重的偏见性后果。这种“部署偏见”源于模型从受控开发环境到复杂社会环境的迁移过程中, 关键的情境变量(Context Variables)被系统性地“蒸发”(Evaporated)了<sup>⑤</sup>。

Selbst 等学者将这种现象精辟地称为“公平性的抽象化陷阱”(Abstraction Traps in Fairness)。具体而言, 部署偏见包含三种子类型: 第一, “框架陷阱”(Framing Trap)——将原本高度嵌入社会结构的问题抽象为可以通过技术优化来解决的“纯工程问题”, 忽略了教育决策中的权力关系和文化语境。例如, 一个在美国常春藤高校数据上开发的学业预警系统若被直接移植到中国的应用型本科院校, 其预警阈值和风险特征向量可能完全不适用——因为两个教育生态的学生群体构成、学术标准体系、文化对“失败”的定义以及可动用的支持资源都存在根本性差异。第二, “涟漪效应陷阱”(Ripple Effect Trap)——未能预见算法决策在社会生态系统中引发的连锁反应。当一个 AI 分班系统将某生分配到“基础班”时, 这不仅影响该生的学业轨迹, 还通过“标签效应”(Labeling Effect)改变了同学、教师乃至该生本人对其能力的认

①Barocas S, Selbst A D. Big Data's Disparate Impact [J]. *California Law Review*, 2016, 104(3): 671-732.

②Dwork C, Hardt M, Pitassi T, et al. Fairness through Awareness [C]. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. New York: ACM, 2012: 214-226.

③Rudin C. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead [J]. *Nature Machine Intelligence*, 2019, 1(5): 206-215.

④Biesta G. Good Education in an Age of Measurement: Ethics, Politics, Democracy [M]. New York: Routledge, 2010.

⑤Selbst A D, Boyd D, Friedler S A, et al. Fairness and Abstraction in Sociotechnical Systems [C]. *Proceedings of the Conference on Fairness, Accountability, and Transparency*. New York: ACM, 2019: 59-68.

知——而这种认知的改变又会以一种自我验证的方式影响该生的后续表现<sup>①</sup>。这正是经典的“皮格马利翁效应”（Pygmalion Effect）在算法时代的回响。第三，“时间漂移陷阱”（Temporal Drift Trap）——模型在训练时使用的数据分布与部署时的现实分布之间随时间推移而出现的偏移。学生群体的构成、学习习惯和技术使用模式每年都在变化，而一个固化在2022年数据上的模型到2026年可能已经严重过时——但由于缺乏持续的公平性监测机制，这种“过时的偏见”往往要造成了大规模的实质性伤害后才会被发现。

#### （四）交互评价偏见：反馈循环与偏见自我强化的闭合回路

AIEd系统并非静态的一次性输出装置，而是在与用户的持续交互中不断进化的动态系统。这种交互本身成为偏见生成和放大的第四重通道——也是最为危险的一重，因为它使偏见获得了“时间维度上的自我复制能力”。

反馈循环偏见（Feedback Loop Bias）是交互偏见中最为凶险的形态，其运作机制可以被概括为一个“偏见—行动—后果—数据—偏见”的闭合回路。第一步，AIEd系统根据其初始预测（可能已含偏见）对某组学生贴上“高风险”标签。第二步，基于该标签，系统或教师减少对这些学生的高阶认知挑战配置，转而提供大量低层次的重复练习。第三步，长期缺乏认知挑战的学生在后续测评中确实表现欠佳。第四步，这些“欠佳”的成绩数据被重新馈入系统，进一步“验证”和加固了最初的偏见性预测。第五步，系统在下一轮迭代中以更高的“置信度”对这组学生做出更为不利的预判<sup>②</sup>。如此反复，形成一个自我强化的恶性循环。关键的洞察在于：在这整个闭合回路中，算法在任何一个时间点上都是“正确”的——因为它的预测确实与实际结果一致。但这种“正确”是一种自我制造的正确，一种通过剥夺机会来验证“没有机会”的预言的荒诞逻辑。

确认偏误效应（Confirmation Bias Effect）是交互偏见的第二重通道。当教师或教学管理者接收到AI系统的预测或建议后，即便该建议实际上含有偏见，人类决策者也倾向于将其视为“数据说话”的客观依据而非需要批判性审查的假设。多项认知心理学研究表明，人类对“来自计算机的判断”具有一种本能的过度信任——即便这种判断实际上并不比随机猜测更准确<sup>③</sup>。这种“算法崇拜”（Algorithm Appreciation）在教育场景中尤其危险：当一位经验丰富的教师对某个AI预测结果感到直觉上的不适时，系统展示的“92.3%置信度”往往会压制教师自身的专业判断——即便教师的直觉是对的。

此外，在个性化学习场景中，AI系统对学生发展前景的“推荐”具有独特的“信息茧房”效应<sup>④</sup>。个性化学习中的推荐算法通过持续追踪学生的点击行为和学习偏好，不断向其推送与既有认知框架高度吻合的内容，逐渐将学生封锁在一个越来越狭窄的“信息牢笼”中。对于一个数学成绩优异但对历史毫无兴趣的学生，算法将以不可逆转的方式不断强化其“数学身份”而持续弱化其人文素养的发展机会——这与教育致力于培养“完整的人”、开阔视野、激发跨领域探索的本质使命构成了根本性的矛盾。更令人忧虑的是，推荐算法的逻辑深处隐藏着一种“效率优先”的价值假设——它将“学生在已擅长领域取得更高分数”视为最优目标，而将“学生在陌生领域遭遇有益的困惑与挑战”视为需要规避的低效路径<sup>⑤</sup>。这种“超级助推”（Hypernudge）的机制，本质上是用工业效率的逻辑来管理人的发展——而人的发展，恰恰是一个需要“低效”的试错、迂回和偶然发现的过程。

综上，算法偏见的全链路生成机制揭示了一个不容回避的核心结论：偏见不是发生在模型训练中某个可被定位和修复的“节点”上的孤立事件，而是弥漫于数据采集→测量构建→模型架构→优化目标→

① Rosenthal R, Jacobson L. *Pygmalion in the Classroom: Teacher Expectation and Pupils' Intellectual Development* [M]. New York: Holt, Rinehart & Winston, 1968.

② Lum K, Isaac W. To Predict and Serve? [J]. *Significance*, 2016, 13(5): 14-19.

③ Logg J M, Minson J A, Moore D A. Algorithm Appreciation: People Prefer Algorithmic to Human Judgment [J]. *Organizational Behavior and Human Decision Processes*, 2019, 151: 90-103.

④ Pariser E. *The Filter Bubble: What the Internet Is Hiding from You* [M]. London: Penguin Books, 2011.

⑤ Yeung K. 'Hypernudge': Big Data as a Mode of Regulation by Design [J]. *Information, Communication & Society*, 2017, 20(1): 118-136.

部署迁移 → 用户交互 → 反馈循环这一完整链路中的系统性过程。正因如此，任何试图仅从模型内部进行“去偏见”修补的策略，都注定只能触及冰山的一角——要真正治理教育 AI 中的算法偏见，必须将治理的视野从模型内部扩展到整个社会-技术生态系统。

本节揭示了算法偏见在教育 AI 系统中的一个根本特征：它不是发生在某个环节的离散故障，而是一种全链路的系统性再生产过程。更为关键的是，教育场景为这种再生产提供了独特的放大机制——由于学生处于发展过程之中，早期的偏见性决策（如低年级阶段的分流建议）将沿着教育管道逐层累积，最终在高等教育入口处形成不可逆的结构性排斥。这正是教育 AI 偏见与刑事司法或信贷领域偏见的本质区别所在：后者是一次性的判决或审批，前者是累积性的发展性伤害。正是这种“偏见的累积性放大效应”，使得一种专门面向教育场景的“发展性公平”概念成为必要。

## 二、从“识别”到“度量”：算法公平性的不可能三角

认识到偏见的存在只是治理的起点。如果说第一节的“全链路考古”回答了“偏见从哪里来”的问题，那么本节要回答的则是一个更为尖锐的操作性追问：我们如何以可量化、可验证、可比较的方式来判断“一个算法是否公平”？这一追问将我们带入了算法公平性（Algorithmic Fairness）的核心理论地带——一个充满数学之美与伦理之痛的交叉领域。

表面上看，“公平”似乎是一个直觉明了的概念：对所有人一视同仁，不偏不倚。然而，当研究者们试图将这一直觉翻译成精确的数学定义时，一个令人震惊的事实浮出水面——“公平”至少有 21 种在数学上互不等价的正式定义<sup>①</sup>。更为惊人的是，这些定义不仅不等价，其中多组还被严格证明为在一般条件下不可能被同时满足。这一发现——被学界称为“公平不可能定理”（Impossibility Theorem of Fairness）——对算法公平性研究构成了根本性的理论挑战，也对教育 AI 的治理实践产生了深远的范式影响。

### （一）群体公平性的核心度量指标体系

在众多公平性定义中，“群体公平性”（Group Fairness）指标家族因其操作性强、易于计算和跨系统可比而获得了最广泛的应用。这一家族的核心思想是：将人口按某一敏感属性（如性别、种族、社会经济地位）划分为不同群体，然后检查模型对不同群体的预测结果是否在某个统计维度上呈现系统性的差异。当前学界被广泛援引的群体公平性指标主要有四个<sup>②</sup>：

人口统计均等（Demographic Parity / Statistical Parity）：这是最直观、也是最古老的公平性定义。它要求模型对不同敏感属性群体的正预测比例应大致相等，而不论各群体的实际基础率如何。用数学语言表达： $P(\hat{Y} = 1|A = a) = P(\hat{Y} = 1|A = b)$ ，其中  $\hat{Y}$  是模型预测， $A$  是敏感属性。在教育场景中，这意味着一个自动分班系统在各种族群体中将学生分配到“高级班”的比例应大致相当。人口统计均等的吸引力在于其简洁性和对“结果平等”的直接承诺。然而，其最大隐患也正在于此：当不同群体的实际能力分布确实存在差异时（这种差异本身可能正是结构性不平等的产物），强制拉平预测比例可能导致大量“误分类”——错误地将不合格学生送入他们尚未准备好的高阶课程，或错误地将合格学生拒之门外<sup>③</sup>。

机会均等（Equal Opportunity）：为回应人口统计均等的上述缺陷，Hardt 等人<sup>④</sup>提出了机会均等指标。它不再要求总体预测比例相等，而是聚焦于“真正合格者”这一子群体，要求在实际为正例的个体中，模型正确识别其为正例的比率（即真阳性率，True Positive Rate）在不同群体之间应当相等： $P(\hat{Y} = 1|Y =$

①Narayanan A. Translation Tutorial: 21 Fairness Definitions and Their Politics [C]. Proceedings of the Conference on Fairness, Accountability and Transparency. 2018.

②Verma S, Rubin J. Fairness Definitions Explained [C]. Proceedings of the International Workshop on Software Fairness. New York: ACM, 2018: 1–7.

③Corbett-Davies S, Goel S. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning [J]. arXiv preprint arXiv:1808.00023, 2018.

④Hardt M, Price E, Srebro N. Equality of Opportunity in Supervised Learning [C]. Advances in Neural Information Processing Systems. 2016: 3315–3323.

$1, A = a) = P(\hat{Y} = 1 | Y = 1, A = b)$ 。在教育场景中,这意味着一个学业预警系统在识别“实际面临困难的学生”时,对不同社会经济背景的学生应当同样敏感。

均等化赔率 (Equalized Odds): 这是机会均等的强化版本。它在要求真阳性率跨群体相等的基础上,进一步要求假阳性率 (False Positive Rate) 在不同群体间也应相等。均等化赔率是最严格的群体公平性标准之一,其实践中的代价是模型的预测准确率往往会显著下降——为了在所有群体间实现统计上的“完美对称”,模型不得不在某些群体上故意“犯错”。

预测值均等 (Predictive Parity / Calibration): 与前三个指标不同,预测值均等关注的不是“谁被预测为正例”,而是“被预测为正例的人中有多少是真阳性”。在教育场景中,这意味着一个 AI 录取系统预测“录取后能成功毕业”的学生中,不同性别或种族群体的实际毕业率应当相当。

## (二) 个体公平性: “相似的人应被相似对待”

与群体公平性从人口统计的宏观视角出发不同,个体公平性 (Individual Fairness) 聚焦于微观层面的“相似性”,其核心原则由 Dwork 等人<sup>①</sup>正式提出:“相似的个体应当被相似地对待” (Similar individuals should be treated similarly)。在数学上,这要求模型满足一种利普希茨条件 (Lipschitz condition)。

个体公平性在教育场景中具有直觉上的强大吸引力:两个在学术能力、学习投入和个人素质上高度相似的学生,仅仅因为一个来自富裕家庭而另一个来自贫困家庭,就被 AI 系统给出截然不同的预测或推荐——这显然是不公平的。然而,个体公平性的致命难题在于“相似性”如何定义。在教育这个高度多维的领域中,什么构成“相似”的学生?“相似性度量” (Distance Metric) 本身就可能是偏见的载体——如果我们使用的相似性度量对来自弱势群体的学生系统性地“放大”差异 (例如,将文化差异误读为能力差异),那么个体公平性非但不能矫正偏见,反而会以“公平”之名将偏见合法化<sup>②</sup>。

## (三) 因果公平性: “如果你不是这个种族,结果会不同吗?”

近年来,因果推断 (Causal Inference) 范式的引入为公平性研究开辟了一条全新的理论路径。因果公平性 (Causal Fairness) 的核心追问不是“不同群体是否被不同对待”,也不是“相似的人是否被相似对待”,而是一个反事实 (Counterfactual) 问题:“如果这个学生的种族/性别/社会经济地位与实际不同,而其他一切条件不变,模型的预测结果是否会改变?”<sup>③</sup>。如果答案是“是”,则该预测在因果意义上是不公平的。

因果公平性为教育 AI 的偏见检测提供了一个极为强大的分析工具。它不仅能够检测“显性歧视”(即模型直接使用了敏感属性),还能够检测“隐性歧视”(即模型通过因果链条间接地受到了敏感属性的影响)。然而,因果公平性在实践中面临巨大的操作困难。首先,构建准确的因果图 (Causal Graph) 需要对教育过程中各变量之间的因果关系有深入的领域知识。其次,反事实推断从根本上是不可验证的——我们永远无法真正观察到“如果这个学生不是女性,她的成绩会怎样”——因此因果公平性的判断在经验层面始终具有假设性质。

## (四) “公平不可能定理”: 教育场景中的深刻裂隙

在梳理了群体公平性、个体公平性与因果公平性三大范式之后,我们抵达了算法公平性理论中最为震撼的核心结论:多组数学家已经从不同角度严格证明——在不同群体的基础率 (Base Rate) 存在差异的一般条件下,上述核心公平性指标不可能被同时满足<sup>④⑤</sup>。

①Dwork C, Hardt M, Pitassi T, et al. Fairness through Awareness [C]. Proceedings of the 3rd Innovations in Theoretical Computer Science Conference. New York: ACM, 2012: 214–226.

②Fleisher W. What's Fair about Individual Fairness? [C]. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. 2021: 480–490.

③Kusner M J, Loftus J, Russell C, Silva R. Counterfactual Fairness [C]. Advances in Neural Information Processing Systems. 2017: 4066–4076.

④Chouldechova A. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments [J]. *Big Data*, 2017, 5(2): 153–163.

⑤Kleinberg J, Mullainathan S, Raghavan M. Inherent Trade-Offs in the Fair Determination of Risk Scores [C]. Proceedings of Innovations in Theoretical Computer Science. 2017: 43.1–43.23.

Chouldechova (2017) 的证明尤为简洁有力：她证明了当两个群体的基础率不等时（即  $P(Y = 1|A = a) \neq P(Y = 1|A = b)$ ），任何非完美的预测器都无法同时满足预测值均等和均等化赔率。Kleinberg 等人 (2017) 则从另一个角度证明了校准性（Calibration）与群体平衡（Balance）之间的不可兼容性。

将这一数学结论置放回教育场景，其含义是极其深刻和令人不安的。在教育领域，不同社会经济背景的学生群体在历史上的学业基础率本身就几乎必然存在差异<sup>①</sup>。面对这种客观存在的基础率差异，“公平不可能定理”告诉我们：试图通过算法优化来同时实现所有维度的“公平”在数学上就是不可能的。这迫使我们直面一个极为艰难的价值抉择：我们愿意在哪个维度上容忍不公，以换取在另一个维度上的公平？这个问题没有技术上的“正确答案”——它本质上是一个需要在特定教育情境下、由所有利益相关者通过充分知情的民主协商来共同界定的价值判断问题。

（五）超越度量：公平作为过程而非状态

“公平不可能定理”的最深层启示，不是“公平是无法实现的”，而是“公平不是一个可以被一劳永逸地‘实现’的终态”。公平不是一个可以通过工程优化达到的静态指标，而是一个需要通过持续的监测、审议、调整 and 问责来不断趋近的动态过程<sup>②</sup>。任何声称“我们的算法已经通过了公平性测试，因此是公平的”的 AIED 供应商都是在进行一种危险的误导。“什么是这个场景下可接受的公平”这一问题的答案，不当应当由算法工程师或教育管理者单方面给出，而应当是一个纳入学生、家长、教师和社区声音的民主协商过程的产物。

表 1: 关键算法公平性度量指标在教育场景中的应用与反思

| 度量指标   | 核心定义                         | 教育应用示例                             | 批判性反思与局限                           |
|--------|------------------------------|------------------------------------|------------------------------------|
| 人口统计均等 | 不同群体获得正预测的比例应相等              | 分班系统中各族裔被分入“高级班”的比例应当              | 可能为追求比例均等而牺牲预测准确性，录取不合格学生或拒绝合格学生   |
| 机会均等   | 真正合格者中，不同群体被正确识别为合格的比率应相等    | 学业预警系统中确实会挂科的不同背景学生被预警的比率应相同       | 仅追溯合格者端的公平，对历史性基础率差异（本身即不平等产物）无法矫正 |
| 均等化赔率  | 真阳性率和假阳性率在不同群体间均应相等          | 辅导干预系统中，需要和不需要干预的学生在不同群体中的判定准确率应相同 | 最严格的公平标准，实践中极难实现；模型预测准确率往往成为牺牲品    |
| 预测值均等  | 被模型预测为正的个体中，实际为正的比例在不同群体间应相等 | AI 录取系统中被录取学生中不同性别学生的学业成功率应当       | 该指标可能导致对历史录取率低的群体采用更高录取门槛，反而加剧不平等  |

本节的核心发现可以被凝缩为一个命题：公平性指标之间的内在矛盾不是技术不成熟的暂时困境，而是“公平”概念本身多维性的不可消解的映射。但本文进一步论证，这一命题在教育场景中获得了独特的理论延伸：教育不是一场一次性的资源分配博弈，而是一个持续数年乃至数十年的发展过程。这意味着教育 AI 的公平性检验不能仅在单一时间截面上进行，还必须追踪其对学生长期发展轨迹的影响。本文将这种面向教育的特殊要求概念化为“发展性公平”（Developmental Equity）：一个教育 AI 系统是否“发展性公平”，不仅取决于它在当下是否对不同群体一视同仁，更取决于它是否为每一个学生——特别是起点最低的学生——保留了充分的发展可能性，而非过早地、不可逆地将其锁定在某条受限的轨道上。“发展性公平”与传统的“静态公平”之间的核心区别在于：前者要求公平性评估必须纳入时间维度和累积效应，后者则仅关注单一决策时点的群体间指标对齐。

①Reardon S F. The Widening Academic Achievement Gap Between the Rich and the Poor [M]// Duncan G J, Murnane R J. Whither Opportunity? New York: Russell Sage Foundation, 2011: 91–116.

②Wong P H. Democratizing Algorithmic Fairness [J]. *Philosophy & Technology*, 2020, 33: 225–244.

### 三、多维审计框架：从技术修复到制度治理

既然单靠算法内部的“去偏见”技术无法从根本上解决教育 AI 中的公平性问题，就必须将治理视角从模型内部扩展到整个 AIED 生态系统。前两节的分析已经充分揭示：偏见的生成是全链路的、系统性的，公平的度量是多元的、内在矛盾的。因此，对偏见的治理也必须是多维的、贯穿全生命周期的。本文构建了一套涵盖数据审计、模型审计、影响评估、合规审查与利益相关者参与五个维度的多维审计框架，力图教育 AI 的公平性治理提供一个既有理论深度又具操作性的系统化方案。

这一框架的理论基石来自两个方面：其一，是 Raji 等学者提出的“端到端内部算法审计”（End-to-End Internal Algorithmic Auditing）范式<sup>①</sup>，该范式强调审计不应仅聚焦于模型输出的某个时间截面，而应贯穿从数据采集到系统退役的完整生命周期；其二，是 Metcalf 等学者关于“算法问责性制度化”的研究<sup>②</sup>，该研究指出，算法审计不能仅仅是一种技术操作，更必须嵌入到组织治理结构之中，成为一种制度化的常规实践。

#### （一）数据审计：追溯偏见的源头

数据审计是整个审计框架的第一道防线，也是成本效益比最高的治理环节——因为在数据层面发现和修正的偏见，其治理成本远低于在模型层面或部署层面发现同样的偏见后再进行补救。数据审计的核心任务是在数据进入模型训练之前，系统地检查数据集的来源可靠性、人口群体覆盖度、标注过程的客观性以及是否存在可能导致歧视的历史偏见编码。

Geburu 等学者在 2021 年发表的标志性论文中提出了“数据集说明书”（Datasheets for Datasets）框架<sup>③</sup>，要求每一个用于 AI 训练的数据集都应当附带一份标准化的“说明书”，涵盖以下七类关键信息：动机（为什么创建这个数据集？）、组成（数据集包含哪些实例？）、采集过程（数据是如何被采集的？）、预处理（数据经过了哪些清洗和转换？）、用途（数据集的推荐用途和不推荐用途是什么？）、分布（数据集将如何被分发？）、维护（谁负责数据集的持续更新？）。这一框架借鉴了电子工程领域的“元器件数据表”（Component Datasheets）理念——正如我们不会在不了解元器件参数规格的情况下设计电路一样，我们也不应当在不了解数据集“参数规格”的情况下训练模型。

在教育场景中，数据审计需要特别关注以下四个维度。第一，人口代表性审计：检查训练数据是否在性别、种族、社会经济地位、城乡分布、残障状态等维度上均匀覆盖目标人群。对于中国教育 AI 产品，尤其需要审查训练数据中东部/西部、城市/农村、汉族/少数民族学生的比例是否与实际服务人群的比例相匹配。第二，标注质量审计：检查数据标注者（Annotators）的构成是否多元，标注标准是否清晰无歧义，标注者之间的一致性（Inter-Annotator Agreement）是否达到可接受水平<sup>④</sup>。在教育数据中，“优秀作文”“高质量课堂互动”等标签的评判本身就高度依赖评判者的文化背景和个人偏好——如果标注者全部来自城市中产阶层，其标注结果将不可避免地反映特定阶层的审美标准和价值取向。第三，历史偏见审计：系统评估数据采集时段内的社会环境是否存在可能被数据“编码”的结构性不平等。例如，如果训练数据涵盖了 2001—2020 年中国高校录取记录，审计人员必须意识到这一时段内存在显著的城乡教育资源分配差异，这些差异已经被“烘焙”进了数据之中。第四，时效性审计：评估数据的“保质期”——一个基于 2018 年学生行为数据训练的模型，在 2026 年的教育环境中是否仍然具有预测效力？特别是新冠疫情对学生学习行为模式的根本性改变（从线下转向线上）是否已被纳入数据更新？缺乏时效性审计的数据集就像一张过期

①Raji I D, Smart A, White R N, et al. Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing [C]. Proceedings of the Conference on Fairness, Accountability, and Transparency. New York: ACM, 2020: 33–44.

②Metcalf J, Moss E, Boyd D. Owning Ethics: Corporate Logics, Silicon Valley, and the Institutionalization of Ethics [J]. *Social Research: An International Quarterly*, 2019, 86(2): 449–476.

③Geburu T, Morgenstern J, Vecchione B, et al. Datasheets for Datasets [J]. *Communications of the ACM*, 2021, 64(12): 86–92.

④Hovy D, Prabhumoye S. Five Sources of Bias in Natural Language Processing [J]. *Language and Linguistics Compass*, 2021, 15(8): e12432.

的地图——它描述的是一个已经不存在的世界。

## （二）模型审计：穿透黑箱的探照灯

如果数据审计是“防患于未然”，那么模型审计则是“查病于已然”——在模型训练完成后，利用一系列系统化的测试方法检验模型的输出是否在不同群体间呈现系统性的差异，以及这种差异是否达到了“不可接受”的程度。

模型审计的技术工具库可以分为三个层次。第一层次是群体差异性检测工具：直接计算第二节中讨论的各项公平性指标（人口统计均等、机会均等、均等化赔率、预测值均等）在不同群体间的数值差异，并判断这些差异是否超过了预定的阈值。例如，美国平等就业机会委员会（EEOC）长期使用的“五分之四规则”（Four-Fifths Rule）规定：如果某一群体的录取率低于最高录取率群体的 80%，则存在“表面上的不利影响”（Prima Facie Adverse Impact）。这一规则的核心逻辑——设定一个量化的“差异容忍阈值”——可以被移植到教育 AI 审计中<sup>①</sup>。

第二层次是可解释 AI（Explainable AI, XAI）工具：当群体差异性检测发现了问题后，下一步就是追问“模型为什么会产生这种差异”。LIME（Local Interpretable Model-Agnostic Explanations）和 SHAP（SHapley Additive exPlanations）是当前最具影响力的两种事后解释方法<sup>②</sup>。LIME 通过在单个预测实例的局部邻域内拟合一个可解释的线性模型来“解释”黑箱模型的行为；SHAP 则基于博弈论中的沙普利值（Shapley Value），将每个特征对预测结果的贡献进行公平分配。在教育场景中，审计人员可以利用 SHAP 分析来检查：在一个学业预警模型中，“邮政编码”这一特征对预警结果的贡献权重是否异常地高？如果是，那么模型很可能在通过邮政编码这一代理变量间接地使用了种族或社会经济地位信息。

第三层次是对抗性审计（Adversarial Auditing）：借鉴信息安全领域的“红队”（Red Team）思维，由独立的审计团队刻意构造旨在“暴露”模型偏见的测试输入<sup>③</sup>。例如，构造两份除了学生姓名（一个典型的“白人名字”，一个典型的“非裔名字”）之外完全相同的虚拟学生档案，然后输入 AI 录取系统观察预测结果是否存在差异。Buolamwini 的“Gender Shades”研究本质上就是一种经典的对抗性审计实践。对于生成式 AI 在教育中的应用，对抗性审计还需要检测“同一提示词在不同人口统计学语境下生成的回复是否存在系统性的质量或语气差异”——这构成了传统审计方法难以覆盖的全新挑战领域。

然而，模型审计本身也面临着深刻的方法论局限。其一是“审计时点”问题：模型审计通常在模型开发完成后、大规模部署之前进行——但模型在部署后会因反馈循环和数据漂移而持续演化，审计时的“公平”不等于部署后的“公平”。其二是“审计覆盖率”问题：在高维特征空间中，穷举所有可能的偏见模式在计算上是不可行的，审计只能覆盖有限的测试场景。其三是“解释的可靠性”问题：LIME 和 SHAP 等事后解释工具本身的忠实度（Faithfulness）——即其解释在多大程度上真正反映了模型的内部决策逻辑——仍然是一个活跃的学术争议。因此，模型审计不能是一次性的“通关考试”，而必须是嵌入系统运行全周期的持续性监测机制。

## （三）影响评估：从控制实验到田野追踪

算法偏见的后果不仅体现在模型输出的统计分布上（这是前两个审计维度所关注的），更深层次地体现在它对真实教育场景中真实学生的实际影响中。一个模型在统计指标上“通过”了所有公平性检测，并不意味着它在部署后不会对某些群体的学生造成实质性的伤害——因为实验室中的统计指标无法完全捕捉复杂社会环境中的涌现效应（Emergent Effects）。因此，影响评估必须走出实验室，进入教育田野。

①Feldman M, Friedler S A, Moeller J, et al. Certifying and Removing Disparate Impact [C]. Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2015: 259–268.

②Lundberg S M, Lee S I. A Unified Approach to Interpreting Model Predictions [C]. Advances in Neural Information Processing Systems. 2017: 4765–4774.

③Sandvig C, Hamilton K, Karahalios K, et al. Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms [C]. Data and Discrimination: Converting Critical Concerns into Productive Inquiry. New York: ACM, 2014: 1–23.

影响评估应当遵循“部署前—部署中—部署后”的三阶段框架。部署前阶段进行“算法影响评估”(Algorithmic Impact Assessment, AIA), 参照环境影响评估(EIA)的成熟方法论, 在AIED系统正式投入使用之前对其可能产生的差异化影响进行系统预测和评估。加拿大联邦政府在2019年率先推出了“算法影响评估工具”(AIA Tool), 要求所有使用自动化决策系统的联邦部门在部署前完成一份标准化的影响评估问卷<sup>①</sup>。部署中阶段进行实时监测: 在AIED系统运行过程中, 持续追踪关键公平性指标的变化轨迹, 设置“预警阈值”——当某一群体的预测准确率突然下降或误报率突然上升时, 系统应自动触发审查流程。部署后阶段进行“回顾性影响评估”: 在系统运行一段时间(例如一个学期或一个学年)后, 通过追踪研究(Longitudinal Study)评估AIED系统对不同群体学生在学业成绩、学习动机、心理健康、教育选择等多维度上的长期差异化影响。

在教育场景中, 影响评估还需要特别关注“隐性影响”(Latent Impact)——那些不会立即显现在量化指标中但会在长期尺度上产生深远后果的影响。例如, 一个个性化推荐系统若持续向女性学生推荐“文科类”内容而向男性学生推荐“理工类”内容, 其短期内的学业成绩指标可能看不出任何“不公平”——但在三到五年的时间尺度上, 它将显著固化性别化的学科选择模式, 重塑学生的职业想象空间, 进而在劳动力市场层面再生产性别不平等。这种“慢性毒药”式的影响, 比“急性中毒”式的显性歧视更加危险, 也更加难以被传统的审计方法所捕获。

影响评估的方法论工具箱应当包括: 随机对照试验(RCT)或准实验设计(如断点回归、双重差分), 以建立因果推断的基础; 焦点小组访谈和深度个案研究, 以捕获统计方法无法触及的主观体验维度; 学生和教师的“算法体验报告”(Algorithm Experience Report), 以系统收集终端用户对AI系统的感知、信任度和不满<sup>②</sup>。特别值得强调的是, 影响评估必须采用交叉性(Intersectionality)视角——不仅检查“女性”和“男性”之间、“农村”和“城市”之间的差异, 还必须检查“农村女性”这一交叉身份群体是否面临着比“农村”或“女性”单一维度更为严重的累积性不利影响<sup>③</sup>。

#### (四) 合规审查: 法律的最后防线

合规审查是审计框架的制度保障层, 它将技术审计和影响评估的发现转化为具有法律约束力的行动要求。在全球范围内, 针对AI系统的法律监管框架正在加速成型, 教育领域作为“高风险应用场景”正日益成为监管关注的焦点。

欧盟《人工智能法案》(AI Act)是全球最具系统性和约束力的AI监管立法。该法案采用“风险分层”(Risk-Based Approach)的监管架构, 将AI应用按风险等级分为四类: 不可接受风险(直接禁止)、高风险(严格监管)、有限风险(透明度义务)和最低风险(无额外义务)。教育领域的AI应用——包括用于录取决策、学业评估、学习路径推荐的系统——被明确划入“高风险”类别<sup>④</sup>。被归类为高风险的AIED系统必须满足一系列严格的合规要求: 建立风险管理系统、确保训练数据的质量与代表性、维护详细的技术文档、实施人类监督机制、达到规定的准确性和鲁棒性标准。这一立法的里程碑意义在于, 它首次在法律层面将“教育AI必须被审计”从一种道德呼吁转化为一种法律义务。

英国政府走了一条不同于欧盟“硬法”路径的“软治理”路线。2022年, 英国中央数字与数据办公室(CDDO)推出了算法透明度记录标准(ATRS), 要求公共部门(包括公立教育机构)在使用算法辅助决

①Government of Canada. Algorithmic Impact Assessment Tool [EB/OL]. (2019-04-12)[2026-04-16]. <https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html>.

②Veale M, Van Kleek M, Binns R. Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making [C]. Proceedings of the CHI Conference on Human Factors in Computing Systems. New York: ACM, 2018: 440.1-440.14.

③Crenshaw K. Demarginalizing the Intersection of Race and Sex [J]. *University of Chicago Legal Forum*, 1989, 1989(1): 139-167.

④European Commission. Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) [R]. Brussels: European Commission, 2021.

策时，公开披露算法的名称和用途、所使用的数据来源、算法逻辑的高层描述、已识别的风险和缓解措施、人类监督的安排。ATRS 的核心理念是“阳光是最好的消毒剂”——通过强制性的信息披露来激发公众监督和学术审查，从而间接推动算法公平性的改善。然而，“软治理”路线的固有风险在于其缺乏强制执行力：如果一个教育机构拒绝填写 ATRS 记录，其面临的后果远不如违反欧盟 AI 法案那样严厉。

中国在 AI 治理方面展现了“政策引导 + 平台治理”的双轨特征。2023 年发布的《生成式人工智能服务管理暂行办法》强调了算法的公平性、透明性与可追责性，要求生成式 AI 服务提供者“采取有效措施防止生成虚假信息，防止生成歧视性内容”。2021 年实施的《互联网信息服务算法推荐管理规定》则对算法推荐系统提出了透明度要求。然而，上述规定在教育 AI 的具体适用性方面仍存在显著的模糊地带：“歧视性内容”的判定标准是什么？“公平性”的度量指标由谁来确定？当不同公平性指标之间发生冲突时（如本文第二节所论证的“不可能定理”），法律如何裁决？这些问题的悬而未决，使得中国教育 AI 的合规审查在实践中面临“有法可依但无章可循”的困境。面向未来，中国亟需出台专门针对教育 AI 系统的公平性审计指南，将第二节中讨论的公平性度量指标体系转化为可操作的法律标准。

#### （五）利益相关者参与：审计的民主化维度

多维审计框架的第五个维度——也是最常被忽视却可能是最为重要的维度——是利益相关者的实质性参与（Meaningful Stakeholder Participation）。前四个维度在本质上都是“专家主导”的活动：由数据科学家检查数据、由审计工程师测试模型、由评估研究者追踪影响、由法律专家审查合规。然而，算法偏见最终伤害的是具体的人——学生、教师、家长——他们的声音如果被排除在审计过程之外，审计本身就有沦为“技术精英自我验证”的风险<sup>①</sup>。

利益相关者参与应当在审计的每一个阶段都得到制度化的保障。在数据审计阶段，来自不同社区的代表应当有权审查训练数据是否包含了对其群体的负面刻板印象或系统性遗漏。在模型审计阶段，教师应当被邀请参与“情境化公平性评估”——即在了解模型的公平性指标后，结合其对具体教育情境的深入理解，判断哪些类型的“不公平”在此时此地是最不可容忍的。在影响评估阶段，学生和家长的体验报告应当被视为与统计数据同等重要的证据来源。在合规审查阶段，公民社会组织和学术机构应当有权对 AIEd 供应商的合规声明进行独立的第三方验证。

这一参与式审计理念的理论基础来自 Arnstein 经典的“公民参与阶梯”（Ladder of Citizen Participation）理论<sup>②</sup>。Arnstein 将公民参与分为八个层次，从“操纵”（Manipulation）和“治疗”（Therapy）等伪参与形式，到“伙伴关系”（Partnership）、“授权”（Delegated Power）和“公民控制”（Citizen Control）等实质性参与形式。当前大多数 AIEd 产品的“用户反馈”机制仅停留在阶梯的最底层——一个“满意/不满意”的按钮不是参与，而是一种参与的幻觉。真正的参与意味着被影响的群体在审计的设计、执行和结论形成过程中拥有实质性的话语权和否决权。

综上所述，多维审计框架的五个维度——数据审计、模型审计、影响评估、合规审查与利益相关者参与——构成了一个相互嵌套、层层递进的治理体系。数据审计确保“原料”的清洁，模型审计检验“产品”的质量，影响评估追踪“使用”的后果，合规审查提供“法律”的保障，利益相关者参与则确保整个审计过程本身是“公正”的。这五个维度缺一不可：没有数据审计的模型审计是“无根之木”；没有影响评估的合规审查是“纸上谈兵”；没有利益相关者参与的技术审计是“精英独白”。唯有将这五个维度整合为一个有机的治理生态系统，才有可能在算法偏见的全链路挑战面前构建起有效的制度性回应。

本节构建的五维审计框架，从操作层面回应了“发展性公平”的制度需求。传统审计仅关注算法输出的群体间统计差异——这是“静态公平”视角下的必要但不充分条件。面向教育的审计框架必须额外纳入

<sup>①</sup>Sloane M, Moss E, Awomolo O, et al. Participation Is Not a Design Fix for Machine Learning [C]. Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization. New York: ACM, 2022: 1–6.

<sup>②</sup>Arnstein S R. A Ladder of Citizen Participation [J]. *Journal of the American Institute of Planners*, 1969, 35(4): 216–224.

影响评估中的“长期发展追踪”维度和利益相关者参与中的“学习者发展权表达”维度，才能真正回应教育场景的发展性特征。本文将这种面向教育的审计范式命名为“发展性审计”（Developmental Audit）：它不仅问“算法对不同群体公平吗”，更问“算法是否限制了某些学生的发展可能性”。

## 四、算法去偏见技术：能力边界与范式局限

在明确了多维审计框架的必要性之后，有必要对当前主流的算法去偏见（Debiasing）技术进行一次系统的“能力体检”——不是为了否定其不可替代的工程价值，而是为了清醒地认识其能力边界、适用条件和深层的范式局限。当前学界和工业界的去偏见技术按照干预时点的不同，大致可以划分为三大类：预处理阶段技术（Pre-processing）、训练中阶段技术（In-processing）和后处理阶段技术（Post-processing）。每一类技术都有其独特的设计理念、操作原理、工程优势和固有局限。本节将对三类技术逐一进行深入解剖，并在此基础上提炼出它们共享的范式性盲区。

### （一）预处理阶段技术：在数据入口处拦截偏见

预处理技术的核心思想是“在偏见进入模型之前就将其清除”——即在模型训练开始之前，对原始训练数据进行修正、转换或增强，使修正后的数据集在统计上更加“公平”，从而使任何在此数据上训练出来的模型都更不容易产生偏见性输出。

预处理技术主要包括以下几种操作策略。第一种是重采样（Resampling）：通过对训练数据中代表性不足的群体进行过采样（Oversampling）——即复制或合成生成更多的少数群体样本——或者对代表性过高的群体进行欠采样（Undersampling）——即随机删除部分多数群体样本——来使数据集在不同群体之间达到数量上的平衡<sup>①</sup>。在教育场景中，如果一个用于预测学生辍学风险的模型的训练数据中，低收入家庭学生的样本仅占总样本的10%（远低于其在实际学生群体中的比例），那么过采样技术可以通过增加低收入家庭学生的样本数量来缓解这种代表性偏差。SMOTE（Synthetic Minority Over-sampling Technique）是最常用的合成过采样方法之一，它通过在少数类样本之间的特征空间中进行插值来生成新的合成样本<sup>②</sup>。

第二种是重新加权（Reweighting）：不改变数据集的构成，而是为不同群体的样本分配不同的权重，使得在模型训练过程中，少数群体的样本被赋予更高的“重要性”，从而抵消其数量不足的影响。第三种是数据变换（Data Transformation）：对数据的特征空间进行数学变换，使变换后的特征与敏感属性之间的相关性降至最低。Zemel等学者提出的“学习公平表征”（Learning Fair Representations, LFR）方法就是这一思路的典型代表<sup>③</sup>，它试图将原始数据映射到一个新的表征空间，使该空间中的表征既尽可能保留了对预测任务有用的信息，又尽可能“遗忘”了与敏感属性相关的信息。

预处理技术的核心优势在于其“模型不可知性”（Model-Agnostic）——由于修正发生在数据层面而非模型层面，它可以与任何类型的下游模型（从简单的逻辑回归到复杂的深度神经网络）组合使用，无需修改模型架构或训练算法。这一特性使得预处理技术在实际部署中具有极高的工程灵活性和可移植性。

然而，预处理技术存在三个根本性的局限。其一，它只能处理“已知的”、“可被定义和测量的”偏见——如果审计人员未能识别出数据中存在某种偏见（例如第一节中讨论的测量偏见），那么预处理技术对此无能为力。其二，数据修正可能引入新的偏差：过采样可能导致模型对少数群体样本的过拟合（Overfitting），欠采样则可能丢失有价值的信息，而合成数据生成可能产生与现实分布不符的“幽灵样本”（Ghost Samples）。其三，也是最深层的局限：预处理技术假设偏见的根源在于数据的不平衡或不清洁，但正如第一节所分析

<sup>①</sup>Kamiran F, Calders T. Data Preprocessing Techniques for Classification without Discrimination [J]. *Knowledge and Information Systems*, 2012, 33(1): 1–33.

<sup>②</sup>Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: Synthetic Minority Over-sampling Technique [J]. *Journal of Artificial Intelligence Research*, 2002, 16: 321–357.

<sup>③</sup>Zemel R, Wu Y, Swersky K, et al. Learning Fair Representations [C]. *Proceedings of the 30th International Conference on Machine Learning*. 2013: 325–333.

的，即便是一个在统计上“完美代表”当前人口的数据集，只要当前社会本身存在结构性不平等，该数据集就天然携带着历史偏见——而这种嵌入社会结构深处的偏见，不是通过调整样本比例就能消除的。

## （二）训练中阶段技术：在优化过程中嵌入公平约束

训练中技术采取了一种与预处理技术根本不同的策略：它不修改数据，而是直接修改模型的训练过程——将公平性要求作为一个约束条件或正则化项嵌入模型的损失函数（Loss Function），使模型在优化预测准确性的同时被迫“兼顾”公平性指标。换言之，训练中技术试图让模型在“学习”的过程中就“学会”公平。

Agarwal 等学者提出的“归约方法”（Reductions Approach）是训练中技术的里程碑之作<sup>①</sup>。该方法的核心创新在于将公平约束分类问题归约为一类成本敏感分类（Cost-Sensitive Classification）问题，使得原本复杂的约束优化问题可以通过迭代调用标准分类算法来求解。这一方法的技术贡献在于它提供了一个统一的理论框架，可以灵活地适配多种不同的公平性定义（如人口统计均等、机会均等等），并给出了关于预测准确性损失的理论保证。

对抗性去偏（Adversarial Debiasing）是训练中技术的另一重要分支，由 Zhang 等学者提出<sup>②</sup>。该方法借鉴了生成对抗网络（GAN）的思想，在模型架构中引入一个“对抗者”（Adversary）组件：主模型（Predictor）试图在做出准确预测的同时“隐藏”敏感属性信息，而对抗者则试图从主模型的预测中“推断”出敏感属性。两者之间的博弈最终推动主模型学会产生“信息不泄露”的预测——即其预测结果不携带关于敏感属性的冗余信息。

正则化方法（Regularization-based Methods）构成了训练中技术的第三条路径。该方法在模型的损失函数中添加一个“公平性惩罚项”（Fairness Penalty），使得模型在训练过程中如果产生了不公平的预测输出就会“受到惩罚”——即其总损失增加。通过调节惩罚项的权重（通常记为超参数  $\lambda$ ），可以在预测准确性和公平性之间进行精细的权衡。

然而，训练中技术面临着三重深刻挑战。第一重挑战是第二节已经论证的“公平不可能定理”：如果不同公平性指标在数学上不可兼容，那么在训练过程中“优化哪个指标”就成为了一个不可回避的价值选择，而这个选择不能由工程师单方面做出。第二重挑战是公平-准确性权衡（Fairness-Accuracy Trade-off）问题：在绝大多数实际场景中，引入公平性约束都会导致模型整体预测准确性的下降<sup>③</sup>。在教育领域，这意味着一个“更公平”的学业预警系统可能在整体上“更不准确”——它可能会漏掉一些真正需要帮助的学生，或者错误地将一些不需要帮助的学生纳入干预范围。这一权衡的存在意味着，即便从纯技术角度出发，去偏见也不是“免费”的——它有真实的精度成本，而这一成本是否值得承担、由谁来承担，本质上是一个需要所有利益相关者共同参与的决策问题。第三重挑战在于计算成本：训练中技术通常需要对标准训练流程进行重大修改，增加了模型开发的复杂性、时间和算力消耗——这对于资源有限的教育机构和中小型 AIED 开发商而言可能构成不可忽视的实际障碍。

## （三）后处理阶段技术：在输出端校准结果

后处理技术采取了最为“务实”的策略：它完全不触碰数据和模型本身，而是在模型训练完成后、输出结果交付给最终用户之前，对模型的预测结果进行校准或调整，以使最终输出满足特定的公平性标准。

Hardt 等学者在其关于“机会均等”的开创性论文中提出了最经典的后处理方法<sup>④</sup>：对不同群体设定

① Agarwal A, Beygelzimer A, Dudík M, et al. A Reductions Approach to Fair Classification [C]. Proceedings of the International Conference on Machine Learning. 2018: 60–69.

② Zhang B H, Lemoine B, Mitchell M. Mitigating Unwanted Biases with Adversarial Learning [C]. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. 2018: 335–340.

③ Corbett-Davies S, Goel S. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning [J]. arXiv preprint arXiv:1808.00023, 2018.

④ Hardt M, Price E, Srebro N. Equality of Opportunity in Supervised Learning [C]. Advances in Neural Information Processing Systems. 2016: 3315–3323.

不同的分类阈值 (Threshold), 使得各群体的真阳性率 (或其他目标指标) 趋于一致。例如, 如果一个学业预警模型在低收入家庭学生群体中的真阳性率为 0.6 (即仅识别出 60% 的真正高风险学生), 而在高收入家庭学生群体中的真阳性率为 0.85, 那么后处理技术可以通过降低低收入群体的分类阈值——使模型对该群体更加“敏感”——来提高其真阳性率至 0.85, 从而实现两个群体之间的机会均等。

校准 (Calibration) 是后处理技术的另一重要方向。Pleiss 等学者提出了在满足校准性约束的前提下实现群间公平的后处理校准方法<sup>①</sup>。该方法的核心思想是确保模型对不同群体的预测概率具有相同的“可信度”——例如, 当模型预测某低收入学生的毕业概率为 70% 时和预测某高收入学生的毕业概率为 70% 时, 两者的实际毕业率应该是相近的。

拒绝选项分类器 (Reject Option Classification) 构成了后处理技术的第三种策略<sup>②</sup>。该方法的核心思想是: 对于那些模型“不太确定”的预测 (即预测概率接近分类阈值的样本), 将有利结果分配给弱势群体成员, 将不利结果分配给优势群体成员, 从而在“边界区域”进行有针对性的再分配。

后处理技术的核心优势在于其极高的实践可操作性: 它不需要访问训练数据或修改模型内部结构——只需要模型的预测输出和敏感属性信息——因此可以被应用于任何已经部署的“黑箱”系统。在教育 AI 的实际治理中, 这一特性意义非凡: 大量已经在学校中使用的 AIED 系统都是由商业公司开发的专有系统, 教育机构无法获得其训练数据或模型代码, 但仍然可以通过后处理技术在输出端对其进行公平性校准。

然而, 后处理技术本质上是一种“事后补救”而非“源头治理”。它没有改变模型内部的偏见性逻辑——模型仍然在用有偏见的方式“思考”, 只是其偏见性的“结论”在输出端被人为矫正了。这意味着后处理方法具有内在的脆弱性: 一旦部署环境发生变化 (如学生群体构成改变) 或公平性要求发生调整 (如从机会均等切换到预测值均等), 先前的校准参数可能立即失效, 需要重新进行调整。更重要的是, 对不同群体设定不同阈值 (即差异化对待) 本身就可能引发法律和伦理上的争议——这种“为了公平而差别化”的操作, 在某些法律框架 (如“形式平等”为核心原则的法律体系) 中可能被视为“反向歧视” (Reverse Discrimination)。

#### (四) 技术去偏的范式局限: 一场永远无法胜利的军备竞赛

在逐一解剖了三类去偏见技术的能力与局限之后, 我们可以提炼出它们共享的、更为根本性的范式性盲区。

第一重范式局限: “偏见是可分离的”假设。三类技术无论在哪个阶段进行干预, 都隐含有一个关键假设: 偏见是一种可以被从数据或模型中“分离”和“移除”的杂质, 就像可以从水中过滤掉泥沙一样。然而, 正如第一节的全链路分析所揭示的, 偏见不是附着在数据表面的“噪声”, 而是深深嵌入数据生成过程、社会结构和文化规范之中的“信号”。在一个深刻不平等的社会中, 数据中的“偏见”和“信息”可能是同一枚硬币的两面——移除偏见可能同时移除了有价值的信息, 保留信息可能同时保留了偏见<sup>③</sup>。Safiya Noble 尖锐地指出, 算法偏见不是一个可以通过更好的代码来“修复”的技术问题, 而是权力结构在数字空间中的系统性再生产。

第二重范式局限: “静态公平”假设。现有去偏见技术几乎无一例外地是在固定的时间截面上、基于固定的数据集、针对固定的公平性指标进行优化的。然而, 教育是一个高度动态的过程: 学生在成长, 教师在调整教学策略, 课程在更新, 社会环境在变化。一个在 2024 年通过了公平性检测的模型, 到 2026 年可能因为学生群体构成的变化而变得不再“公平”。更为棘手的是, 去偏见操作本身可能触发新的反馈循环: 如果一个去偏后的模型改变了对某些群体学生的资源分配, 这种改变将影响这些学生的后续表现, 而

<sup>①</sup>Pleiss G, Raghavan M, Wu F, et al. On Fairness and Calibration [C]. Advances in Neural Information Processing Systems. 2017: 5680–5689.

<sup>②</sup>Kamiran F, Karim A, Zhang X. Decision Theory for Discrimination-Aware Classification [C]. Proceedings of the IEEE International Conference on Data Mining. 2012: 924–929.

<sup>③</sup>Noble S U. Algorithms of Oppression: How Search Engines Reinforce Racism [M]. New York: NYU Press, 2018.

后续表现数据又将被用于更新模型——形成一个新的、可能完全不可预测的动态循环<sup>①</sup>。Liu 等学者的研究揭示，在某些条件下，旨在“帮助”弱势群体的公平性干预，由于其对长期激励结构的改变，反而可能在长期内损害该群体的利益——这一“公平性干预的延迟效应”构成了去偏见技术最为反直觉的理论发现之一。

第三重范式局限：“技术中立”假设。去偏见技术的整个范式建立在一个隐含的前提之上：技术干预是“中立的”——工程师只是在执行一个由外界给定的“公平性目标”，技术本身不参与对“什么是公平”的界定。然而，恰恰是在将抽象的公平性概念翻译成具体的数学公式的过程中，大量的价值判断被悄无声息地做出了。选择优化人口统计均等还是机会均等——这不是一个技术决策，而是一个深刻的伦理-政治决策：前者隐含着“结果平等”的价值取向，后者则隐含着“机会平等”的价值取向<sup>②</sup>。当工程师在代码中写下一个特定的公平性约束时，他/她实际上是在为整个受算法影响的群体“代选”了一种价值立场——而这种代选在绝大多数情况下既未经过受影响者的知情同意，也未经过公共审议。

第四重范式局限：“生成式 AI 的去偏困境”。随着 ChatGPT、Claude、Gemini 等大语言模型（LLM）在教育领域的爆发式渗透，去偏见技术面临着一场前所未有的范式跃迁挑战。传统的去偏见技术——无论是预处理、训练中还是后处理——都是针对“分类/预测型”模型设计的，即模型输出一个离散的标签（如“高风险/低风险”）或一个连续的概率值。然而，生成式 AI 的输出是连续的自然语言文本，其中的偏见嵌入方式远比一个预测标签更加隐蔽、微妙和多维。当一个 LLM 为学生生成课程建议或论文反馈时，其偏见可能表现为微妙的语气差异、隐含的期望差异、建议的深度差异或使用的例证的文化偏向——这些差异几乎不可能被任何现有的量化公平性指标所捕获。如何为生成式文本“定义”和“度量”公平，目前在学术上仍然是一个基本未解的问题<sup>③</sup>。

综言之，算法去偏见技术作为一个工程工具箱，在特定的、定义清晰的应用场景中具有不可否认的实用价值：它们能够在给定公平性目标下显著降低模型输出的群体间差异，在一些教育 AI 应用中取得了可测量的改善效果。然而，将去偏见技术视为教育算法偏见的“解决方案”——而非“缓解手段之一”——则是一种危险的范畴错误。偏见的根源深嵌于社会结构之中，而技术只能在数据和模型的表层进行修补。试图仅凭技术手段来“解决”一个本质上是社会-政治的问题，就像试图仅凭退烧药来“治愈”一场由基础免疫缺陷引发的感染——它可以暂时缓解症状，但无法根除病因，甚至可能通过制造“已经治愈”的假象而延误真正的治疗。正是这一认识，构成了下一节从“修复模型”向“人机协同治理”进行范式转型的理论基础。

本节的“范式体检”揭示了一个对“教育发展性公平”具有特殊意义的发现：三类去偏技术无一能够应对偏见的“累积性放大效应”。它们都是在单一决策时点上、针对单一模型进行的一次性干预，而教育中的偏见是沿着“幼儿园—小学—初中—高中—大学”这条管道逐层累积的。一个在小学阶段被去偏后的推荐系统，可能在表面上实现了群体间的“公平”——但如果它的上游（学前教育数据）和下游（初中分流系统）都充斥着未经纠正的偏见，那么这种“点状公平”在整个教育管道中的实际效果将趋近于零。“发展性公平”要求的不是“点状去偏”，而是“管道贯通式治理”——这正是技术手段力所不及、而制度治理必须介入的核心原因。

①Liu L T, Dean S, Rolf E, et al. Delayed Impact of Fair Machine Learning [C]. Proceedings of the International Conference on Machine Learning. 2018: 3150–3158.

②Friedler S A, Scheidegger C, Venkatasubramanian S. On the (im)possibility of Fairness [J]. arXiv preprint arXiv:1609.07236, 2016.

③Liang P, Bommasani R, Lee T, et al. Holistic Evaluation of Language Models [J]. *Annals of the New York Academy of Sciences*, 2023, 1525(1): 140–146.

## 五、从“修复模型”到“人机协同治理”：范式转型与全球比较

前四节的分析已经充分揭示了教育 AI 算法偏见问题的全貌：偏见的生成是全链路的（第一节），公平的度量是内在矛盾的（第二节），审计的治理必须是多维的（第三节），而技术去偏见的能力是有根本性边界的（第四节）。这些发现共同指向一个不可回避的结论：教育 AI 的公平性问题不能被还原为一个纯粹的技术优化问题，它需要一场从“修复模型”到“治理系统”的范式转型——而这正是本节所要论述的核心主题。

### （一）对技术至上主义的批判：解构“算法修复叙事”

技术至上主义（Techno-Solutionism）是一种根深蒂固的意识形态倾向，它笃信每一个社会问题——无论其根源多么深厚、多么复杂——都可以通过更好的技术来解决<sup>①</sup>。Evgeny Morozov 将这种倾向精辟地概括为“对每一个问题都点击‘解决’”的冲动。在 AIED 领域，技术至上主义的典型话语是：“算法有偏见？那就用更好的算法去消除偏见。”这一话语的危险性在于：它将一个本质上是社会-政治的问题伪装成了一个纯技术问题，从而将解决方案的制定权从公众的民主审议中转移到了技术精英的封闭实验室中。

技术至上主义在教育 AI 领域的表现形式至少有以下三种。第一种是“更多数据叙事”：认为偏见是数据不足或数据质量差的结果，只要收集更多、更全面、更高质量的数据，偏见就会自然消失。这一叙事忽略了第一节已经论证的核心观点——即便是一个在统计上完美代表当前人口的数据集，只要当前社会本身存在结构性不平等，该数据集就天然携带着历史偏见。第二种是“更好算法叙事”：认为当前算法之所以产生偏见，是因为去偏见技术还不够成熟，只要算法继续进步，终将找到一种既完美准确又绝对公平的解决方案。这一叙事被第二节的“公平不可能定理”从数学上证伪——不同公平性指标之间的内在矛盾不是技术进步可以消解的，而是公平概念本身的多维性和特定社会情境的不可化约性的反映。第三种是“透明度万能叙事”：认为只要算法足够透明——即所有人都能理解算法如何工作——偏见问题就会得到解决。然而，透明度虽然是必要条件，但远非充分条件：即便一个算法的偏见被完全可视化了，如果缺乏有效的问责机制和纠错权力，知道偏见的存在并不等于能够消除偏见<sup>②</sup>。

对技术至上主义的批判并不意味着技术是无用的或有害的。相反，本节的论点是：技术是治理生态系统中不可或缺的组成部分，但它必须被嵌入一个更广阔的社会-制度框架之中才能发挥其最佳效用。将技术从“解决方案”重新定位为“工具之一”，不是降低了技术的地位，而是将其放在了一个更加诚实、更加可持续的位置上。

### （二）“人在环中”的必要性与制度设计

教育 AI 系统中的任何高风险决策——录取、分班、评分、预警、资源分配——都必须贯彻“人在环中”（Human-in-the-Loop, HITL）原则：算法可以作为决策的“辅助者”和“建议者”，但最终的判断权必须保留在具有专业知识、情境感知能力和伦理判断力的人类教育工作者手中。

HITL 原则的理论合法性来自两个方面。第一，认识论维度：教育决策的本质特征是高度情境化的（Context-Dependent）。一个学生的“学业风险”不能被简化为一组量化指标的线性组合——它涉及家庭环境的突然变化、同伴关系的微妙波动、心理健康的隐性困扰等大量无法被数据完全捕捉的因素。教师——特别是与学生建立了长期信任关系的班主任——所掌握的“隐性知识”（Tacit Knowledge）是任何算法都无法替代的认知资源<sup>③</sup>。第二，伦理论维度：将关乎人的命运的决策完全委托给一个无法承担道德责任的算法系统，在伦理上是不可接受的。算法不能“为”其决策后果承担责任——它没有意识、没有意图、没有道德能力。如果一个学生因为算法的错误预测而被错误地分流到低期望的教育轨道，谁来为这一后果负

①Morozov E. To Save Everything, Click Here: The Folly of Technological Solutionism [M]. New York: PublicAffairs, 2013.

②Ananny M, Crawford K. Seeing without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability [J]. *New Media & Society*, 2018, 20(3): 973-989.

③Polanyi M. The Tacit Dimension [M]. Chicago: University of Chicago Press, 1966.

责？算法本身不能被追责，因此必须有一个能够被追责的人类决策者处于决策链条之中。

然而，仅仅要求“有一个人在环中”是不够的——HITL 原则能否真正发挥作用，取决于“人”是否具备足够的能力和权力来对算法的建议进行实质性的审查和否决。研究表明，“自动化偏见”（Automation Bias）——即人类决策者过度信赖自动化系统的输出而忽视自身判断的倾向——是 HITL 模式的最大威胁<sup>①</sup>。在教育场景中，一个工作负荷沉重、缺乏算法素养的教师，面对一个声称基于“大数据”和“先进算法”的 AI 系统给出的“科学”建议，可能倾向于不加批判地接受——这使得“人在环中”蜕变为“人在旁观”（Human-on-the-Loop），HITL 的保护功能形同虚设。

因此，HITL 原则的有效实施至少需要三个制度性前提条件。第一，算法素养培训的制度化：教师和教育管理者必须接受系统化的培训，使其能够理解 AI 系统的基本工作原理、识别常见的偏见模式、评估 AI 建议的可靠性，并在必要时做出独立于 AI 建议的专业判断。第二，否决权的制度化保障：教师和学生必须被赋予对 AI 决策提出质疑和申诉的正式权利——不仅仅是“可以不同意”，而是“有权利和渠道推翻 AI 的建议并获得替代性的人类判断”。第三，时间和资源的保障：HITL 模式意味着教师需要花费额外的时间来审查 AI 的建议、做出独立判断、记录决策理由——如果教师的工作负荷已经过重，HITL 就不可能从一项原则变成一种实践。

### （三）教育算法治理的组织架构：从个体负责到机构问责

仅仅依靠教师个体的算法素养和批判性判断来抵御算法偏见，是一种不切实际的期待——这等于将一个系统性问题的解决方案寄托在个体的道德勇气之上。有效的教育 AI 治理需要在机构层面建立常设性的治理架构。

本文建议，每一个大规模采用 AIED 系统的教育机构（如省级教育行政部门、大型学区、重点高校）都应设立“教育算法治理委员会”（Education Algorithm Governance Committee, EAGC）。EAGC 应当是一个跨学科的常设机构，其成员至少应包括：教育学者（提供教育专业判断）、计算机科学家（提供技术审计能力）、法律专家（提供合规审查）、伦理学者（提供价值分析）、家长代表（提供社区视角）和学生代表（提供受影响者视角）。EAGC 的核心职能应包括：审查拟采购 AIED 系统的公平性评估报告、监督已部署系统的持续公平性监测、受理和调查来自教师、学生和家长的算法偏见投诉、定期发布机构层面的“算法影响年度报告”<sup>②</sup>。

EAGC 模式的理论基础来自 Selbst 等学者提出的“社会技术系统”（Sociotechnical Systems）框架。该框架的核心论点是：算法公平性不能被抽象地、脱离具体社会情境地讨论——一个算法是否“公平”，不仅取决于其数学属性，更取决于它被嵌入的组织结构、被使用的方式、以及它所影响的具体人群。因此，算法治理不能仅仅是“技术治理”，更必须是“组织治理”——它需要在组织层面建立起发现偏见、评估偏见、纠正偏见和问责偏见的完整制度链条。

### （四）全球比较与中国方案

当前全球 AIED 治理呈现三条清晰的路径分化，每一条路径都深深扎根于其特定的政治文化和法律传统之中<sup>③</sup>。

欧盟路径以法律硬约束为主导。以《人工智能法案》为核心的欧盟方案，代表了全球最为系统化和强制化的 AI 治理模式。在教育领域，欧盟将用于录取、评估和分流的 AI 系统划入“高风险”类别，要求开发者在系统投入使用前必须完成风险评估、建立质量管理体系、确保人类监督机制的有效运行。这一模式的核心优势在于其法律约束力——不遵守者将面临高额罚款（最高可达全球营业额的 6% 或 3000 万欧元），

<sup>①</sup>Parasuraman R, Manzey D H. Complacency and Bias in Human Use of Automation: An Attentional Integration [J]. *Human Factors*, 2010, 52(3): 381–410.

<sup>②</sup>Selbst A D, Boyd D, Friedler S A, et al. Fairness and Abstraction in Sociotechnical Systems [C]. *Proceedings of the Conference on Fairness, Accountability, and Transparency*. New York: ACM, 2019: 59–68.

<sup>③</sup>Jobin A, Ienca M, Vayena E. The Global Landscape of AI Ethics Guidelines [J]. *Nature Machine Intelligence*, 2019, 1(9): 389–399.

这为 AI 公平性治理提供了“有牙齿的”执行力。然而，其潜在风险在于过度监管可能抑制 AI 在教育领域的创新应用，以及统一的监管标准可能无法充分适应不同教育情境的多样性需求。

英国路径以行业标准与透明度驱动为特征。与欧盟的“硬法”方案不同，英国更倾向于通过部门级的软性框架来推动 AI 治理。Alan Turing 研究所发布的一系列 AI 伦理指南和公平性工具包为教育机构提供了实践性的指导，而非强制性的法律要求<sup>①</sup>。算法透明度记录标准（ATRS）则要求公共部门在使用算法辅助决策时进行信息披露。这一模式的优势在于灵活性和适应性——它允许不同教育机构根据自身的特点和需求采取差异化的治理策略。然而，其风险在于缺乏强制执行力：在一个安全性投入成本高昂而短期收益不明显的市场中，自愿性的伦理承诺可能沦为“伦理洗白”（Ethics Washing）的工具<sup>②</sup>。

中国路径呈现“政策引导 + 平台治理”的双轨特征。2023 年发布的《生成式人工智能服务管理暂行办法》和 2021 年实施的《互联网信息服务算法推荐管理规定》构成了中国 AI 治理的政策框架。在教育领域，中国的治理路径具有两个显著特征：其一，政策推动力度强——一旦决策层关注到算法公平性问题，其治理方案的推行速度可能远快于欧美；其二，平台企业在治理中扮演重要角色——中国的教育 AI 市场高度集中于少数头部企业（如科大讯飞、好未来等），这使得通过平台治理来推动行业规范成为一条可行路径。然而，如第三节合规审查部分所分析的，当前中国教育 AI 治理面临“有法可依但无章可循”的困境——关键概念（如“歧视性内容”“公平性”）的操作化定义仍然缺失。

然而，中国教育 AI 场景对现有算法公平理论提出了三项独特的修正要求。第一，城乡二元结构重塑了公平基线（Fairness Baseline）：中国教育中最深层的不平等不是种族或性别维度的（尽管这些维度同样存在），而是城乡维度的——城市重点学校与农村薄弱学校之间在数据丰度、硬件条件和教师能力方面的鸿沟，使得任何以“全国均值”为基线的公平性指标都可能遮蔽最需要被看见的不公平。第二，应试体制决定了偏见的独特传导机制：在高考作为“终极分流器”的体制下，教育 AI 的偏见不仅在横向上分化群体，更在纵向上沿着“小升初—中考—高考”的管道被层层放大——这种“偏见的管道化累积”是中国教育场景对“发展性公平”概念的最直接经验支撑。第三，平台化学习的中国特征要求差异化治理：中国教育 AI 市场的高度平台集中（作业帮、学而思、猿辅导等少数平台覆盖数亿用户）意味着，少数平台的算法偏见将对整个国家的教育公平产生系统性影响——这一规模效应在欧美分散化的市场结构中并不存在，因而需要一种平台级的、而非产品级的治理框架。

面向中国高等教育的 AIED 偏见治理，本文提出以下五项本土化建议。第一，建立教育 AI 产品准入审查制度：参照医疗器械的分级管理模式，对教育 AI 产品按其决策影响力进行分级，对高风险产品（如用于录取和分流的系统）要求在投入使用前提交公平性评估报告。第二，推动教育数据治理的法制化建设：制定教育数据保护专项法规，明确数据采集的边界、数据使用的限制以及数据主体（学生和家長）的知情权和异议权。第三，在教师教育体系中系统嵌入“算法素养”（Algorithm Literacy）课程：不仅要培训教师使用 AI 工具的技术能力，更要培养教师对 AI 建议进行批判性审查的专业能力——包括识别偏见模式、评估预测可靠性、做出独立专业判断的能力。第四，建立学生算法权利保障机制：确保每一位受到 AI 辅助决策影响的学生都有权获得关于决策依据的可理解解释，并有权对自己认为不公平的 AI 决策提出正式申诉。第五，推动对教育 AI 系统的独立第三方审计：由学术机构或经认证的独立审计机构对市场上主流的教育 AI 产品进行定期的公平性审计，审计结果向社会公开发布，以社会监督弥补行政监管的不足。

#### （五）教育 AI 公平性治理的哲学基础：重新定义“教育中的公平”

在提出具体的治理方案之后，有必要退后一步，追问一个更为根本的问题：在教育 AI 的语境中，“公

①The Alan Turing Institute. Understanding Artificial Intelligence Ethics and Safety: A Guide for the Responsible Design and Implementation of AI Systems in the Public Sector [R]. London: The Alan Turing Institute, 2019.

②Wagner B. Ethics as an Escape from Regulation: From Ethics-Washing to Ethics-Shopping [C]. Being Profiled: Cogitas Ergo Sum. Amsterdam: Amsterdam University Press, 2019: 84-89.

平”究竟意味着什么？如前文第二节所论证的，“公平”不是一个单一的、自明的概念，而是一个多维的、内在矛盾的价值集合体。不同的公平性定义反映了不同的正义理论：人口统计均等对应的是结果平等（Equality of Outcome）的正义观，机会均等对应的是机会平等（Equality of Opportunity）的正义观，而个体公平性则更接近于一种反歧视（Anti-Discrimination）的正义观。这些正义观之间的紧张关系，在教育领域尤为突出。

Rawls 的“差异原则”（Difference Principle）为教育 AI 的公平性讨论提供了一个有价值的哲学参照点<sup>①</sup>。差异原则认为，社会中的不平等只有在对最弱势成员最有利的条件下才是正当的。将这一原则应用于教育 AI，意味着一个“公平”的 AIEd 系统不应仅追求在不同群体之间实现某种统计上的对称——它更应该被设计为优先改善最弱势学生群体的教育境况。这一哲学立场对去偏见技术的实践具有深刻的指导意义：它暗示，在某些情况下，“公平”可能恰恰要求“不对称”的对待——即对弱势群体给予更多的关注和更优质的支持——而不是追求所有群体的指标对齐。

Sen 的“能力方法”（Capability Approach）则从另一个角度挑战了现有公平性框架的局限性<sup>②</sup>。Sen 认为，衡量公平的标准不应是人们实际获得的“结果”（如考试成绩），而应是人们拥有的“实质自由”——即他们能够选择和实现有理由珍视的生活方式的能力。将能力方法应用于教育 AI 评估，意味着一个“公平”的 AIEd 系统不仅要关注学生的学业成绩是否改善，更要关注系统是否扩展了学生的选择空间、增强了学生的自主学习能力、提升了学生参与知识建构的能动性。一个提高了所有学生考试成绩但同时收窄了学生探索兴趣 a 和自主选择空间的 AIEd 系统，从能力方法的视角来看，可能并不“公平”——即便它在所有现有的量化公平性指标上都表现优异。

这些哲学反思的实践意义在于：它们提醒我们，教育 AI 的公平性治理不能被简化为一个“选择正确的数学指标”的技术问题。在数学指标的背后，是关于“我们要什么样的教育”“我们要什么样的社会”的根本性价值选择。任何声称可以“客观地”“科学地”解决教育公平问题的 AI 系统，都在有意或无意地遮蔽这些选择的政治性质。真正负责任的教育 AI 治理，应当使这些价值选择变得可见、可争论、可修正——而不是将它们隐藏在算法的黑箱之中。

## 六、结论

本文围绕“教育发展性公平”（Developmental Equity in Education）这一核心概念，以“认识—治理—反思”三重递进逻辑为主线，论证了教育 AI 中的算法偏见不是一般意义上的技术偏差，而是教育不平等的自动化再生产机制。经过五节逐层深入的论证，本文得出以下核心结论。

### （一）研究发现的理论贡献

第一，偏见的“全链路考古”揭示了偏见生成的系统性本质。第一节分析表明，算法偏见并非一个可以被精确定位在数据集或模型参数某个“节点”上的离散性缺陷，而是一个弥漫于“数据采集—特征工程—模型训练—系统部署—用户交互”全生命周期的系统性过程。数据层面的历史偏见（如基于历史录取数据训练的模型必然继承过去的歧视模式）、模型层面的归纳偏见（如深度学习模型对多数群体模式的优先学习）、部署层面的反馈循环偏见（如预测导致的资源再分配塑造了被预测对象的未来表现）以及交互层面的确认偏见（如教师对 AI 建议的选择性信赖），共同构成了一个自我强化的偏见生态系统。这一发现的理论意义在于，它从根本上否定了“偏见可以通过修复某个环节来消除”的简单化思路，要求我们采取全生命周期的系统治理视角。

第二，“公平不可能定理”的教育实证化揭示了公平概念的内在矛盾性。第二节分析表明，群体公平性（Demographic Parity）、机会均等（Equalized Odds）、预测值均等（Predictive Parity）等主流公平性

<sup>①</sup>Rawls J. A Theory of Justice [M]. Revised edition. Cambridge, MA: Belknap Press of Harvard University Press, 1999.

<sup>②</sup>Sen A. Development as Freedom [M]. New York: Anchor Books, 1999.

指标之间存在数学上不可调和的矛盾——在基础率（Base Rate）不同的群体之间，不可能同时满足所有公平性标准。这一发现的深远意义在于，它使我们不得不承认：“公平”不是一个具有唯一正确解的技术优化问题，而是一个需要在特定社会-教育情境中通过多方协商来“界定”——而不是“计算”——的价值命题。教育领域恰恰是这种紧张关系最为尖锐的场域之一：对一个被学业预警系统“标记”的学生而言，“公平”可能意味着获得更多的关注和支持；但对一个因为来自“高风险群体”而被系统化地低期望对待的学生而言，同样的“关注”可能构成另一种形式的歧视。

第三，多维审计框架为教育 AI 治理提供了系统化的操作方案。第三节构建的涵盖数据审计、模型审计、影响评估、合规审查与利益相关者参与五个维度的审计框架，不仅是一个理论贡献，更是一个具有实践操作性的治理工具。该框架的核心创新在于将审计从一种“技术检测”活动扩展为一种“制度治理”实践——它不仅关注算法输出是否在统计上“公平”，更关注审计过程本身是否公正、透明、可问责和民主参与。

第四，对技术去偏见的“范式体检”划定了技术治理的能力边界。第四节对预处理、训练中和后处理三类去偏见技术的系统性解剖揭示了四重根本性的范式局限：“偏见可分离”假设忽略了偏见与信息不可分解性；“静态公平”假设未能应对教育环境的高度动态性；“技术中立”假设遮蔽了将公平概念数学化过程中不可避免的价值选择；生成式 AI 的爆发式应用更使传统去偏范式面临根本性的适用性危机。这些发现的核心价值在于，它们为技术工具在教育 AI 治理中的角色进行了诚实而精确的定位——技术是不可或缺或“缓解手段”，但绝非“终极解决方案”。

第五，从“修复模型”到“人机协同治理”的范式转型为教育 AI 偏见治理提供了新的理论框架和制度设计方案。第五节对技术至上主义的批判揭示了“算法修复叙事”的意识形态本质，“人在环中”原则的制度化论证为人类教育工作者的不可替代性提供了认识论和伦理论的双重辩护，“教育算法治理委员会”的组织架构设计为机构层面的问责机制提供了可操作的蓝图，而 Rawls 差异原则和 Sen 能力方法的引入则为教育 AI 的“公平”概念提供了更为深刻和丰富的哲学基础。

## （二）对中国教育 AI 治理的政策启示

基于上述理论发现，本文对中国教育 AI 治理提出以下政策启示。

第一，尽快建立教育 AI 产品的分级分类管理制度。当前中国的 AI 监管框架在教育领域的适用性仍然存在显著的模糊地带。本文建议参照医疗器械管理模式，将教育 AI 产品按照其决策影响力分为低风险（如智能题库推荐）、中风险（如个性化学习路径规划）和高风险（如录取辅助决策、学业预警与分流建议）三个等级，对高风险产品要求在投入使用前提交独立的第三方公平性评估报告并通过审查。

第二，将“算法素养”纳入教师教育的核心课程体系。如本文第五节所论证的，“人在环中”原则的有效运行依赖于教师具备对 AI 建议进行批判性审查的专业能力。然而，当前中国教师教育体系中几乎完全缺乏这一维度的课程设置。本文建议：在师范院校的培养方案中增设“教育 AI 伦理与算法素养”必修课程；在在职教师继续教育中设立“算法素养”专项培训模块；在教师资格认证标准中纳入“理解 AI 辅助决策的基本原理和局限性”的能力要求。

第三，建立学生和家长的“算法知情权”和“算法申诉权”保障机制。任何受到教育 AI 辅助决策影响的学生及其监护人，都应当有权获得关于决策过程和依据的可理解解释，有权拒绝完全由算法做出的高风险决策，有权对自己认为不公平的 AI 决策向机构内的独立申诉机制提出正式申诉。这一机制的建立不仅是对学生权利的保障，也是对教育 AI 系统持续改进的重要反馈来源。

第四，推动建立全国性的教育 AI 公平性基准测试数据集和评估标准。当前中国缺乏专门针对教育 AI 系统的公平性评估标准和基准测试体系。本文建议由教育部牵头，联合信息技术标准化机构和高校研究团队，开发一套涵盖不同教育场景（录取、评估、推荐、预警）、不同敏感属性维度（性别、城乡、民族、残疾状态、社会经济地位）和不同公平性指标（群体公平、个体公平、因果公平）的综合性评估框架和基准

数据集，为教育 AI 产品的公平性审计提供统一的度量标尺。

### （三）研究的局限与未来方向

本文的研究存在以下局限。第一，在实证层面，由于目前公开可得的中国教育 AI 系统公平性评估数据极为有限，本文的论证主要依赖国际文献和理论推演，未能对中国教育 AI 系统的实际偏见状况进行基于原始数据的实证检验。未来研究应着力开展针对中国主流教育 AI 产品（如自适应学习平台、智能阅卷系统、学业预警系统）的大规模实证审计，以提供本土化的偏见证据。第二，在理论层面，本文虽然讨论了生成式 AI 对传统去偏范式的挑战，但未能深入系统地构建一套专门适用于生成式教育 AI 的公平性度量与治理框架。鉴于大语言模型在教育领域的渗透速度，这一方向具有极高的学术紧迫性。第三，在制度层面，本文提出的“教育算法治理委员会”等组织设计方案尚未经过实证检验，其在中国现有教育行政体制中的可行性和有效性仍需通过试点实验来验证。

面向未来，教育 AI 偏见治理的终极目标不应是制造一个“无偏见”的完美算法——这既在数学上被证明为不可能（公平不可能定理），也在哲学上不可取——而应当是将“教育发展性公平”从一个学术概念转化为一套可操作的制度实践。这意味着构建一个使偏见能够被持续发现、被公开审议、被民主问责、被迭代修正的制度生态系统。在这个生态系统中，技术审计工具、法律监管框架、组织治理架构和民主参与机制共同发挥作用；在这个生态系统中，公平性评估不再局限于单一时间截面上的群体间指标对齐，而是纳入了时间维度、累积效应和发展权保障的全过程追踪。“教育发展性公平”概念的核心贡献在于它回答了一个此前悬而未决的问题：为什么教育 AI 的公平性不能照搬刑事司法或金融信贷领域的既有框架？因为教育面对的不是被“判断”的静态客体，而是正在成长中的发展性主体——他们的未来不应被算法的历史性偏见所预先封锁。唯有将这一洞察嵌入技术设计、制度规范和社会治理的每一个层面，教育 AI 才不会沦为社会不平等的“数字放大器”，而有可能成为促进教育公平和人的全面发展的“正义基础设施”。

引用本文 丁勇. 教育人工智能中的算法偏见：生成机制与治理框架 [J]. 人工智能教育研究, 2026, 2(1):84–108. <https://doi.org/10.6914/aiesc.020103>.

Cite This Article Yong DING.(2026). Algorithmic Bias in AI-Powered Education: Mechanisms and Governance. *Artificial Intelligence Education Studies (Réngōng zhìnéng jiàoyù yánjiū)*, 2(1):84–108. <https://doi.org/10.6914/aiesc.020103>

## Algorithmic Bias in AI-Powered Education: Mechanisms and Governance

Yong DING

Weifang University of Science and Technology

**Abstract** With the rapid expansion of artificial intelligence in education (AIEd), algorithmic bias has become a critical challenge for educational equity. Taking AIEd as the analytical context, this paper examines the generative mechanisms and governance strategies of algorithmic bias from a socio-technical systems perspective. The study argues that algorithmic bias in educational AI is not merely a technical defect but a cumulative reproduction of structural inequalities across the entire pipeline of data collection, model training, system deployment, and human-machine interaction. A full-pipeline analytical framework is developed to identify four sources of bias: data bias, model bias, deployment bias, and interaction bias. Building on the literature of algorithmic fairness, the paper further analyzes the tensions among group fairness, individual fairness, and causal fairness. The study proposes the concept of developmental equity, emphasizing that fairness in educational AI should be understood as a dynamic process oriented toward learners’ long-term development rather than a static statistical balance. Based on this perspective, a multidimensional governance framework—including data

auditing, model auditing, impact assessment, and institutional regulation—is proposed to support responsible development and governance of AI in education.

**Keywords** AI in Education; Algorithmic Bias; Developmental Equity; Algorithmic Fairness; AIEd Governance

〔责任编辑：黄欣 Email: cpcl@cpcl.hk〕